

Assessing and Teaching Radiotherapy Contouring



Dr Simon Lewis Duke

Medical Education Centre, School of Medicine

This thesis is submitted for the degree of
Doctor of Philosophy

March 2021

Abstract

Advanced radiotherapy techniques such as image-guided adaptive brachytherapy for cervical cancer improve local tumour control and reduce treatment toxicity. This benefit is critically dependent on radiotherapy targeting or “contouring” by oncologists. Numerous studies have shown considerable inter-observer contouring variation across all tumour sites, often measured in centimetres, suggesting that current methods of teaching contouring are ineffective. Moreover, assessing contouring competency is currently a subjective, time-consuming and onerous process.

The aim of this programme of research is to investigate the assessment and teaching of radiotherapy contouring within an educational design research framework. The thesis reviews the limitations and challenges of current strategies to improve radiotherapy contouring and how insights from the educational literature such as cognitive load theory, deliberate practice theory, and best practices in assessment and feedback can inform and improve contouring assessment and teaching. Real-world data from two studies of online assessment and education for radiotherapy contouring, within an international clinical trial of advanced radiotherapy techniques for locally advanced cervical cancer, were analysed to substantiate the limitations of current approaches within a clinical trial setting.

The thesis describes a novel low-fidelity radiotherapy contouring simulation tool developed to address some of the issues identified in the clinical studies. A detailed useability study was carried out in a small group of oncologists, which also yielded interesting insights into their clinical reasoning and self-regulation processes. The simulation was then used in three pilot studies of different types of learners (trainees and experts) and programmes (one-off workshops and longitudinal programmes) to explore its acceptability, useability and effectiveness.

The thesis concludes by discussing possible approaches for the next iteration of software development and educational research, which could lead to meaningful change in the teaching and assessment of radiotherapy contouring.

Acknowledgements

Any doctoral research owes a debt of gratitude to many, and this is no exception. In addition to securing the initial funding and creating a vision for the programme of research, my co-principal supervisor Li-Tee Tan has been a constant supporter, guide, constructive critic and inspiration. I will benefit from her mentorship throughout the rest of my career. I am grateful to Gill Doody for spearheading the Nottingham supervisory team and always providing thought-provoking, helpful and pragmatic guidance. Thank you to Rakesh Patel for introducing me to the community of academic medical education, and for inspiring me with his constantly high standards of educational theory and practice. Thank you also to Heather Wharrad for providing knowledge, guidance and engaging conversation in the technology-enhanced learning sphere.

I am incredibly grateful to Addenbrooke's Charitable Trust and its donors for funding this research, and to Robin Crawford, Pippa Corrie, and Hugo Ford at Cambridge University Hospitals for enabling me to secure this funding; also to Rob Glynne-Jones for listening to and refining my research plans as well as providing additional funding. It has been an honour and a great pleasure to become part of the EMBRACE 'family' over the last four years. I am especially grateful for the warm welcome, encouragement, stimulating collaboration and mentorship of Richard Pötter, Kari Tanderup, Remi Nout, Alina Sturdza, Max Schmid, and Kathrin Kirchheiner.

Thank you to all the study participants who gave their consent and time, to which I hope I have been able to do justice.

I have benefitted from many other advisers, such as Raj Jena who provided advice on data analysis, Hatem Helal who helped me grapple with and learn to love MATLAB, and Charlotte Coles for her advice on my poster presentation prior to the ESTRO award. Adam Dorling - we could not have made Mini-Contour without you - thank you so much. Thank you also to Michelle Arora for being a Cambridge 'Med-Ed guru' always ready to share her wisdom. I am thankful for the unwavering support and proof-reading skills of my wonderful parents and in-laws Lewis, Jenny, Dylan and Trix. There are too many others to list them all, but I hope I can honour their contribution by assisting others in a similarly wholehearted and selfless way.

I am grateful to God for the all wonderful opportunities that this research project has brought, and finally would like to pay tribute to the enthusiastic, loving support and understanding of my wife Heather and my wonderful children.

Research outputs

Publications

- Implementing an online radiotherapy quality assurance programme with supporting continuous medical education – report from the EMBRACE-II evaluation of cervix cancer IMRT contouring. **Duke SL**, Tan LT, Jensen NBK, Rumpold T, de Leeuw A, Kirisits C, et al. *Radiotherapy and Oncology* 2020 Jul; 147: 22-29.
- Education and training for image-guided adaptive brachytherapy for cervix cancer—The (GEC)-ESTRO/EMBRACE perspective. Tan LT, Tanderup K, Kirisits C ... **Duke SL [15th/29 authors]** ... Van Limbergen E, Haie-Meder C, Potter, R. *Brachytherapy* 2020 Aug 16;S1538-4721(20)30130-6. doi: 10.1016/j.brachy.2020.06.012.
- Image-guided Adaptive Radiotherapy in Cervical Cancer. Tan LT, Tanderup K, Kirisits C, de Leeuw A, Nout R, **Duke S** ... Jürgenliemk-Schulz I, Lindegaard JC, Pötter R. *Seminars in Radiation Oncology* 2019 Jul; 29(3):284-298. doi: 10.1016/j.semradonc.2019.02.010.
- Challenges for Quality Assurance of Target Volume Delineation in Clinical Trials. Chang ATY, Tan LT, **Duke S**, Ng W-T. *Frontiers in Oncology*. 2017;7.

Prizes

- Poster prize (international*): Implementing a novel online education programme to support RTQA – the EMBRACE-II experience. S.L. Duke, N.B.K. Jensen, T. Rumpold, R.A. Nout, A.A.C. De Leeuw, R.C. Pötter, J.C. Lindegaard, I.M. Jürgenliemk-Schulz, K. Tanderup, L.T. Tan - *ESTRO congress, Barcelona 2018 - abstract PO-0809*. **Received ctRO award (€1000 & free open-access publication) for best clinical poster in the ‘young investigator’ category, one of 6 awards in a field of over 1600 posters and e-posters.*
- Poster prize (regional): S. Duke, R. Potter, D. Gregory, L. Tan. Targeting high cervix cancer cure rates with low toxicity, worldwide: quality assurance of brachytherapy in the EMBRACE-II trial. *NIHR Eastern Cancer Research Network Conference 2019 - 1st prize.*
- Oral presentation prize (local): Improving radiotherapy accuracy: designing the building blocks for an effective online education programme. *University of Nottingham Sue Watson PhD Presentation - awarded 2nd prize by faculty & also the audience prize.*
- Oral presentation prize (local): Research ‘Impact Statement’ presentation prize. *Nottingham Faculty of Medicine Postgraduate Research Forum June 2018.*

- European Society for Radiation Oncology (ESTRO) scholarship to attend Association for Medical Education in Europe 2018 Annual Scientific Meeting.

Other presentations

- Oral: Enabling deliberate practice: using a web-based simulation of radiotherapy targeting to provide rapid individualised feedback to Clinical Oncology trainees. S. Duke, L. Tan, G. Eminowicz, E. Park, H. Wharrad, R. Patel, G. Doody. *Association for the Study of Medical Education Annual Scientific Meeting, Glasgow 2019.*
- Oral: A systematic analysis of delineation performance seen in EMBRACE-II brachytherapy quality assurance. S. Duke, R. Pötter, A. Sturdza, M. Schmid, T. Rumpold, U. Mahantshetty, N. Nesvacil, A. de Leeuw, C. Kirisits, K. Tanderup, R. Nout, J. Lindegaard, I. Jürgenliemk-Schulz, L.T.Tan. *European Society of Radiotherapy and Oncology (ESTRO) congress, Milan 2019 - abstract OC-0176.*

Table of Contents

1	Introduction to radiotherapy as a cancer treatment	1
1.1	Radiotherapy as a cancer treatment: key concepts	1
1.2	The radiotherapy process and “volume delineation” or “contouring”	2
1.3	Selected advances in radiotherapy	6
1.4	Radiotherapy contouring – a post-graduate educational challenge	10
1.5	Summary	12
2	Radiotherapy contouring variation - a critical review of the literature	13
2.1	Introduction	13
2.2	Methods and materials	14
2.3	Variation in contouring – magnitude, causes and consequences	14
2.4	Minimising contouring variability : strategies	20
2.5	Discussion	25
2.6	Summary	27
3	Identifying and applying relevant educational theory	28
3.1	Introduction	28
3.2	Methods and materials	29
3.3	Results	31
3.4	Simulation	33
3.5	Cognitive load theory	39
3.6	Deliberate practice theory	45
3.7	Assessment	51
3.8	Feedback	57
3.9	Summary	61
4	Methodology	62
4.1	Overarching research questions	62
4.2	Epistemological considerations	62
4.3	Educational design research	64
4.4	Programme of research & study cohorts	68
5	EMBRACE-II EBRT accreditation: online education to support radiotherapy quality assurance	72
5.1	Introduction	72
5.2	Methods and materials	75
5.3	Results	79
5.4	Discussion	87

5.5 Conclusion.....	91
6 EMBRACE-II brachytherapy quality assurance: contouring assessment	93
6.1 Introduction	93
6.2 Methods and materials	97
6.3 Results.....	103
6.4 Discussion	115
6.5 Conclusion.....	119
7 Initial development of Mini-Contour: a low-fidelity radiotherapy contouring simulation.....	121
7.1 Review of ‘analysis & exploration’ findings.....	121
7.2 Design & construction	123
7.3 Foundations, initial specification & early testing	123
7.4 Next iteration.....	130
8 Mini-Contour: usability study.....	135
8.1 Introduction	135
8.2 Methods and materials	137
8.3 Results.....	142
8.4 Discussion	158
8.5 Conclusion.....	164
9 Teaching contouring using Mini-Contour: three pilot studies	166
9.1 Introduction	166
9.2 Methods and materials	168
9.3 Results - Exercise creation.....	174
9.4 Results - UK workshops	175
9.5 Results - international trainee longitudinal programme	188
9.6 Results - EMBRACE-II contouring workshop.....	199
9.7 Discussion	209
9.8 Conclusion.....	219
10 Conclusion	220
References	226
APPENDICES	249

Table of Figures

Figure 1-1 – The radiotherapy process.....	3
Figure 1-2 - Target volumes in radiotherapy - a simplified illustration.....	4
Figure 1-3 - The process of radiotherapy "contouring" or "target volume delineation"	6
Figure 1-4 - Using advanced radiotherapy techniques to shape the radiation dose.....	7
Figure 1-5 - Brachytherapy dose drops sharply as the distance from the source increases.....	9
Figure 1-6 - X-ray-guided brachytherapy compared to adaptive MRI-guided brachytherapy	10
Figure 2-1 - Patient survival in the TROG 02.02 study in head and neck cancer, by compliance to radiotherapy protocol and tumour control probability (TCP) classification.....	19
Figure 2-2 - Flow diagram of the radiotherapy quality assurance (RTQA) process for a benchmark case.....	24
Figure 3-1 - Hermeneutic framework for literature reviews.	30
Figure 3-2 - Thematic map of domains identified from a search of medical education textbooks, arranged by relevance to simulated assessment and teaching of practical skills	32
Figure 3-3 - Effective instructional design features in medical simulation.	36
Figure 3-4 - The Atkinson-Shiffrin model of human memory & information processing.....	39
Figure 3-5 - Estimated accumulated practice hours of four cohorts of Berlin violinists in Ericsson et al.'s seminal work on deliberate practice.....	45
Figure 3-6 - Schematic illustration of deliberate practice as a journey towards expertise with increasingly sophisticated performance representations.	46
Figure 3-7 - Practice hours as a function of age in Mcnamara et al.'s study attempting to replicate Ericsson et al.'s methods	49
Figure 3-8 - Miller's pyramid.....	51
Figure 3-9 - Model of feedback and associated variables.....	59
Figure 4-1 - The iterative process of educational design research.....	65
Figure 4-2 - Stokes' classification of research endeavours.....	65
Figure 4-3 - McKenney & Reeves' generic model for educational design research.	66
Figure 5-1 - The Addenbrooke's Contouring Tool (ACT)	73
Figure 5-2 - Risk-adaptive elective lymph node CTV (CTV-E) in EMBRACE-II protocol.....	74
Figure 5-3 - Derivation of the Jaccard conformity index	78
Figure 5-4 - Average scores per ROI for EMBRACE-II EBRT delineation accreditation case	80
Figure 5-5 - Examples of participant errors and assigned scores for various regions of interest.	81
Figure 5-6 - Engagement of EMBRACE-II clinicians with optional CME content.....	86

Figure 6-1 - Target concepts in image-guided adaptive brachytherapy for cervix cancer.	95
Figure 6-2 - Impact of contouring on the reported dose to 90% of the HR-CTV ("D90").....	96
Figure 6-3 - Case 1 Clinical and MRI findings at diagnosis	99
Figure 6-4 - Case 1 clinical and MRI findings at brachytherapy.....	100
Figure 6-5 - Case 2 clinical and MRI findings at diagnosis.....	101
Figure 6-6 - Case 2 clinical and MRI findings at brachytherapy.....	101
Figure 6-7 - Number of regions of interest (ROIs) failed per 1st attempt submission.	104
Figure 6-8 - Average scores (1st attempt) per region of interest for the EMBRACE-II Brachytherapy contouring quality assurance	104
Figure 6-9- Example contouring errors and associated scores for the GTV _{res} and HR-CTV. Yellow = gold standard contour; red = participant GTV _{res} ; magenta = participant HR-CTV.	109
Figure 6-10 - Example contouring errors and associated scores for the IR-CTV. Yellow = gold standard contour; green = participant IR-CTV.	110
Figure 6-11 - Example contouring errors and associated scores for selected organs at risk.....	111
Figure 6-12 Boxplot of Jaccard Conformity Index per target ROI in cases 1 & 2	112
Figure 6-13 Boxplot of Jaccard Conformity Index per organ at risk ROI in cases 1 & 2	112
Figure 7-1 - Olszewski & Wolbrink's framework for virtual simulation development.....	123
Figure 7-2 - Low-fidelity contouring simulation developed by Dr Tan in 2014.....	124
Figure 7-3 - Mini-Contour prototype: user view on accessing a learning exercise.....	125
Figure 7-4 -Mini-Contour prototype: the user contours the target and presses the submit button.	126
Figure 7-5 - Mini-Contour prototype: the reference contour is revealed.....	126
Figure 7-6 - EMBRACE 2018 workshop: participant contours from an external beam radiotherapy exercise.	129
Figure 7-7 - EMBRACE 2018 workshop: delineation of the residual gross tumour volume (GTV _{res}) at time of brachytherapy.	129
Figure 7-8 - First sketch of the 'learning zone' concept	131
Figure 7-9 - User flow through a learning exercise in Mini-Contour version 1.0	134
Figure 8-1 - Usability questions and methods to address them.....	136
Figure 8-2 - Usability study procedures flowchart.....	138
Figure 8-3 - Example of transcribed speech and user actions for the Mini-Contour usability study	140
Figure 8-4 - Usability issue frequency coded by activity	143
Figure 8-5 - Usability issues by exercise number	144
Figure 8-6 - Mini-Contour exercise menu page.....	144
Figure 8-7 - Examples of reduced browser space for Mini-Contour.....	145

Figure 8-8 - Illustration of the difference between contrast and windowing	146
Figure 8-9 - A user's difficulty relating the learning zone written feedback to the displayed areas	147
Figure 8-10 - A user reviewing learning zones and debating the stringency of automated assessment: screen shot and associated transcript	153
Figure 8-11 - An example of a user evaluating learning zone feedback.....	154
Figure 8-12 - Thematic map of users' comments in Mini-Contour usability study	157
Figure 8-13 - Learning zones: separating feedback from assessment	161
Figure 9-1 - Flowchart of Mini-Contour exercise creation and associated study endpoints.....	168
Figure 9-2 - Flowchart of UK trainee workshop study procedures	169
Figure 9-3 - Flowchart of international trainee longitudinal programme study procedures.....	171
Figure 9-4 - Flowchart of EMBRACE 2019 contouring workshop study procedures	172
Figure 9-5 - Average time taken for each stage of the Mini-Contour trainee EBRT exercise creation process	174
Figure 9-6 - An EMBRACE 2019 contouring workshop brachytherapy Mini-Contour exercise	175
Figure 9-7 - UK trainee participant demographics.....	176
Figure 9-8 - Time taken to contour per exercise for the UK trainee cohort	177
Figure 9-9 - Average time - measured in minutes - taken per case for each UK trainee.....	177
Figure 9-10 - UK trainees' average pre-workshop confidence per cervix cancer EBRT region of interest.....	180
Figure 9-11 - Bubble chart comparing UK trainees' confidence in contouring the pelvic lymph node CTV with their stage of training.	181
Figure 9-12 - Performance of UK trainees by Jaccard conformity Index (JCI) across repeated exercise themes.....	183
Figure 9-13 - UK trainees' performance for include learning zones during live workshop.....	183
Figure 9-14 - UK trainees' performance for exclude learning zones during live workshop	183
Figure 9-15 - Performance & confidence on the inclusion zones & related regions of interest for trainees who completed the follow-up exercises	184
Figure 9-16 - Performance & confidence on the exclusion zones & related regions of interest for trainees who completed the follow-up exercises.....	185
Figure 9-17 - Analysis of the sensitivity of selected learning zones to variations in stringency in the UK trainees cohort.....	186
Figure 9-18 - UK trainee contours that failed the ' exclude bladder ' learning zone that would have passed that particular learning zone if 2% overlap was allowed	187
Figure 9-19 - International trainee longitudinal participant demographics	189

Figure 9-20 - Number of participants completing each session of longitudinal Mini-Contour training programme	189
Figure 9-21 - Time taken to contour per exercise for longitudinal trainee contouring programme	190
Figure 9-22 - Time taken to review feedback per exercise for longitudinal trainee contouring programme.....	191
Figure 9-23 - Histogram of time taken for international trainees to review feedback.....	191
Figure 9-24 - A trainee contour (in blue) resulting in a comment regarding overly stringent assessment.....	193
Figure 9-25 - Confidence at baseline and after the first workshop for all trainees in the longitudinal programme.....	194
Figure 9-26 - Confidence for 4 regions over time for the 11 most engaged trainees in the longitudinal programme.....	195
Figure 9-27 - Conformity indices for repeated exercises over the course of the international longitudinal programme - all participants.....	196
Figure 9-28 - International trainees' performance on 'include' (A) and 'exclude' (B) learning zones repeated over the course of the programme - for all participants.....	198
Figure 9-29 - EMBRACE group 2019 annual meeting workshop: participant experience and training in cervix cancer IMRT	200
Figure 9-30 - Time taken for EMBRACE-II clinicians to contour per exercise	201
Figure 9-31 - Histograms of time taken for self-directed EMBRACE participants to submit a contour (A) and review feedback (B).....	201
Figure 9-32 - Boxplots showing System Usability Scale scores across all three cohorts	202
Figure 9-33 - Contours of two EMBRACE-II clinicians (blue) who disagreed with a HR-CTV learning zone.....	204
Figure 9-34 - Contours of two EMBRACE-II clinicians (blue) who disagreed with the 'exclude cervix' learning zone for the residual GTV.....	204
Figure 9-35 - EMBRACE-II: reported confidence (pre- and post-workshop) for target volumes tested.....	206
Figure 9-36 - Conformity for exercises repeated across all three cohorts.....	207

Table of Tables

Table 3-1 - Kirkpatrick's hierarchy of training evaluation applied to medical education.....	38
Table 3-2 - Key constructs in cognitive load theory and their definitions	39
Table 3-3 - Application of cognitive load theory to instructional design in radiotherapy contouring	44
Table 3-4 - Features of mastery learning	47
Table 3-5 - Key terms in assessment relevant to this thesis	52
Table 3-6 - Sources of assessment validity with potential application to radiotherapy contouring assessments	54
Table 3-7 - Characteristics of feedback relevant to practical skills training	57
Table 4-1 - Contrasting the objectivist and interpretivist scientific paradigms.....	63
Table 4-2 - Aims of the EMBRACE-II study.....	69
Table 5-1- Risk groups for defining the elective lymph node clinical target volume in EMBRACE-II.....	73
Table 5-2 - Common and/or clinically significant errors seen in first submissions for EMBRACE-II IMRT benchmark case	82
Table 5-3 - Comparison of pass rates for 56 participants using various JCI cut-offs with expert assessments	84
Table 6-1 - Regions of interest and target concepts for cervical cancer image-guided adaptive brachytherapy	94
Table 6-2 - Assessment scale for EMBRACE-II brachytherapy regions of interest.....	102
Table 6-3 Pairwise comparison of failure rates per clinician and per case, listed by region of interest (ROI). A pass on each ROI was defined as a score of ≥ 6	105
Table 6-4 - Common and/or clinically important delineation errors seen in the EMBRACE-II brachytherapy target volume contouring quality assurance, per ROI.....	107
Table 6-5 - Common and/or clinically important delineation errors seen in the EMBRACE-II brachytherapy organ at risk contouring quality assurance, per ROI.....	108
Table 6-6 - Automatic pass classification rates using Jaccard Conformity Index (JCI) vs manual assessment for selected cutoffs - per ROI	114
Table 6-7 - Common patterns of regression for cervix cancer brachytherapy target contouring	117
Table 7-1 - Potential advantages of low-fidelity software for contouring assessment and teaching	122
Table 8-1 - Usability study endpoints	137

Table 8-2- Stages of qualitative data analysis.....	140
Table 8-3 - Severity grading of usability issues.....	141
Table 8-4 - Punctuation denoting features or framing of participant quotes.....	141
Table 8-5 - Usability study participant characteristics.....	142
Table 9-1 - Mini-contour pilot study primary and secondary endpoints	167
Table 9-2 - UK Trainees' pre-workshop average confidence and correlation of individual confidence with ranked performance on the first relevant learning exercise	181
Table 9-3 - Location and timing of international trainees participating in longitudinal study ..	188
Table 9-4 - Relationship between confidence and performance for international trainees	195
Table 9-5 - Mean conformity indices at baseline, 1st repeat and follow-up for the UK and international trainees	197
Table 9-6 - Conformity per cohort for the three repeated exercises	207
Table 9-7 - Mean learning zone score for the three repeated exercises across the 3 pilot studies	207
Table 9-8 Rates of learning zone errors in the EMBRACE-II workshop compared to rates of the same error in the accreditation exercises.....	209

List of abbreviations & terms

CTV	Clinical target volume. Encompasses potential <i>microscopic</i> local and/or regional tumour spread
EBRT	External Beam Radiotherapy
ESTRO	The European Society of Radiation Oncology
(radiotherapy image) Fusion	The process of combining imaging modalities for a patient, superimposing one upon the other(s) to aid identification of radiotherapy targets and/or organs at risk
(radiotherapy) Fraction	A discrete radiotherapy treatment. Fractions are mostly delivered on different days, but sometimes more than one fraction is given in a day
GTV	Gross tumour volume in radiotherapy contouring. Denotes visible or “macroscopic” tumour
GTV_{res}	The residual gross tumour volume at the time of cervix cancer brachytherapy
Gy	Gray - the unit of radiation dose
HR-CTV	High-risk clinical target volume for cervix cancer brachytherapy. Denotes the residual visible (on scanning or clinical examination) tumour plus areas of fibrosis. The brachytherapy dose is often prescribed to this volume
IGABT	Image-guided adaptive brachytherapy
IGRT	Image-guided radiotherapy
IR-CTV	Intermediate risk clinical target volume for cervix cancer brachytherapy. Denotes the pattern of visible tumour spread at diagnosis superimposed on the anatomy at the time of brachytherapy
ITV	Internal target volume. An expansion of the clinical target volume based on patterns of predicted or measured motion
IMRT	Intensity-modulated radiotherapy
JCI	Jaccard conformity Index
MRI	Magnetic resonance imaging

OaR	Organ at risk. Normal tissues that can be damaged by radiation causing physiological dysfunction and affecting quality of life
PI	Principal investigator (of a clinical trial)
PTV	Planning target volume. The volume to which external beam radiotherapy dose is planned and prescribed. Includes an expansion to allow for inaccuracies in treatment delivery
RCR	Royal College of Radiologists. The professional body of Clinical Oncologists (who deliver radiotherapy) in the UK
ROI	Region of interest in radiotherapy contouring
RTQA	Radiotherapy quality assurance
TCP	Tumour control probability

1 Introduction to radiotherapy as a cancer treatment

Radiotherapy is the main curative treatment modality for several cancer types and is indicated in at least 50% of patients receiving cancer treatment (Barton et al., 2014). In the last 20 years there have been significant advances in the delivery of radiotherapy which have translated into higher cure rates and reduced side effects for cancer patients (Ahmad et al., 2012).

For those who are unfamiliar with radiotherapy this introduction will provide a brief overview of radiotherapy and its use in cancer treatment. I will then highlight some of the major recent innovations in radiotherapy treatment, including advanced external beam radiotherapy and image-guided brachytherapy for cervix cancer, which are the main foci of the radiotherapy aspects of this thesis. The educational challenges that these advances in radiotherapy bring with them are discussed with reference to the targeting of radiotherapy.

1.1 Radiotherapy as a cancer treatment: key concepts

Radiotherapy is the use of radiation to treat illness, most commonly malignant tumours, or “cancers”ⁱ (Cancer Research UK, 2016).

Radiotherapy can be delivered either:

- From outside the body – most commonly using photons, which can be thought of as high-energy X-rays. This is generally termed “**external beam radiotherapy**” (EBRT). Other methods of delivery include the use of electrons, protons and even carbon ions.
- From inside the body, most commonly “**brachytherapy**” – inserting or implanting radioactive material such as Iridium-192. Other methods include “isotope therapy” – using radioactive material ingested and absorbed, or directly infused, into the blood stream.

ⁱ Radiotherapy is also used to treat some benign tumours such as vestibular schwannoma and is occasionally used to treat benign, non-neoplastic conditions

Radiotherapy is an essential element of curative treatment of cancers of the breast, cervix, head and neck, prostate and lung (Jaffray and Gospodarowicz, 2015). In cervix, head and neck, prostate and lung cancers (and other less common cancers) radiotherapy is often used as the primary curative treatment modality. In breast cancer it is most commonly used as an “adjuvant” treatment after surgery, to reduce the chances of local recurrence of the cancer.

Radiation kills cancer cells via DNA damage, but also causes collateral damage to healthy organs or tissues that are near the tumour. In the case of cervix cancer these include the bladder, bowel and femur; in radiotherapy these organs are termed “normal tissues” or “**organs at risk**” (OaRs) (Marks et al., 2010). Giving a high radiation dose to the tumour in pursuit of cancer cure may cause damage to the surrounding organs at risk. This damage can result in severe side effects - in the case of cervix cancer treatment these include urinary or faecal incontinence, bleeding, femoral fracture, fistulae, chronic pelvic infection or even death.

The delivery of radiotherapy is therefore a careful balance of giving a high radiation dose to tumour and areas of potential spread whilst limiting the dose, and therefore the damage, to the organs at risk.

1.2 The radiotherapy process and “volume delineation” or “contouring”

The steps in radiotherapy process are outlined in Figure 1-1 below:

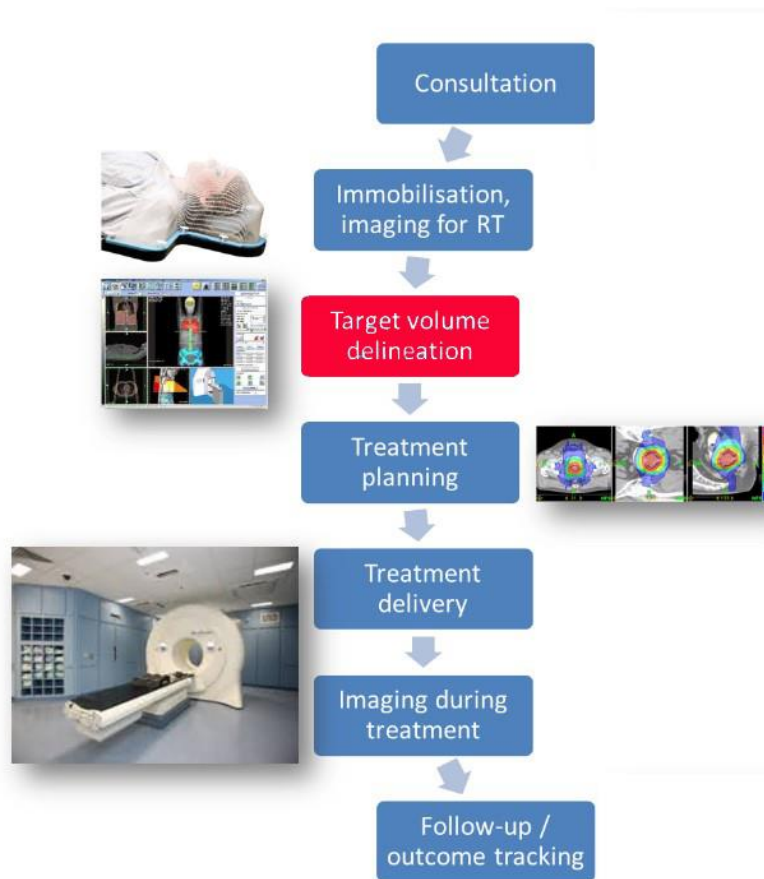


Figure 1-1 – The radiotherapy process. Image courtesy of Dr Raj Jena. Target volume ‘delineation’ is also known as ‘contouring’, ‘segmentation’ and ‘outlining’.

This involves multiple staff groups including radiographers, physicists and radiation oncologistsⁱⁱ. Radiotherapy “contouring”, also known as “delineation”, “outlining” and “segmentation”, forms the basis for the subsequent radiotherapy treatment design and delivery. It is a pivotal step in the radiotherapy process.

The process of radiotherapy contouring entails the clinician identifying and delineating on medical images (most commonly a CT scan) the areas to which radiation dose should be delivered. This in itself is a multi-faceted process, outlined by the International Commission on Radiation Units reports 50 & 62 (ICRU, 1993, ICRU, 1999). Target volumes are identified based on the location of the “gross” (visible) tumour and patterns of potential spread and motion:

ⁱⁱ In the UK, ‘Clinical Oncologists’ are the physicians who oversee the planning and delivery of radiotherapy (they also prescribe chemotherapy). In much of the rest of the world physicians with this responsibility are termed ‘Radiation Oncologists’ and do not administer chemotherapy.

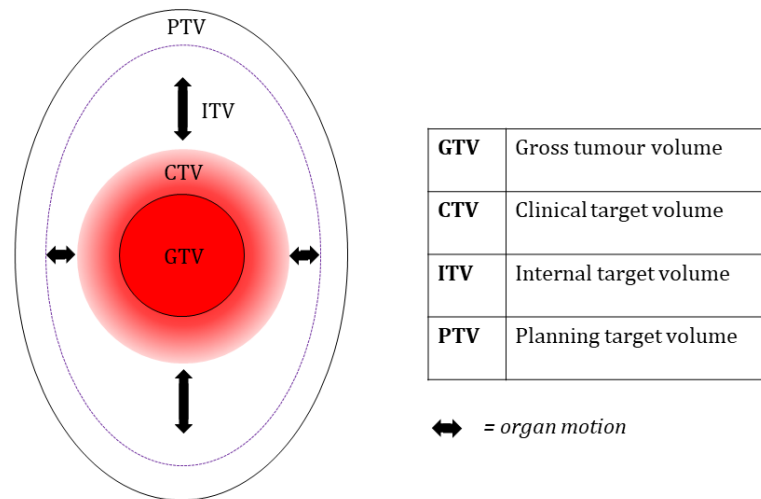
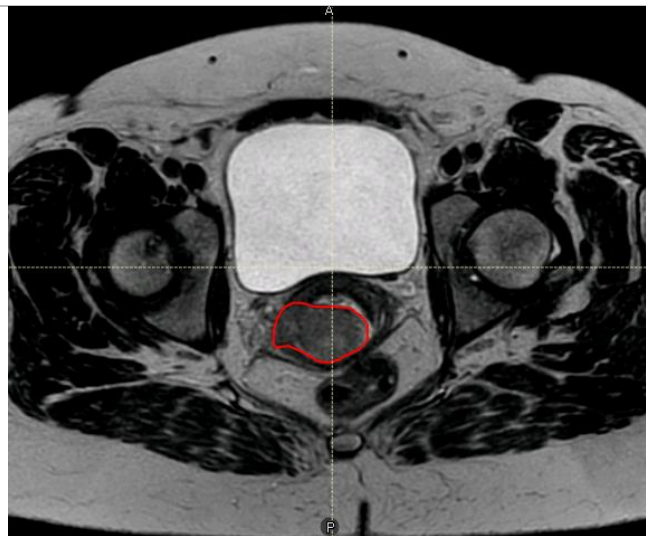


Figure 1-2 - Target volumes in radiotherapy - a simplified illustration. Adapted from Burnet et al. (Burnet et al., 2004) & ICRU (ICRU, 1993, ICRU, 1999)

These target volumes are explained and illustrated for a cervix cancer case below:

Areas of visible (“macroscopic”) tumour are identified, using information from imaging (MRI in this case) and / or clinical examination.

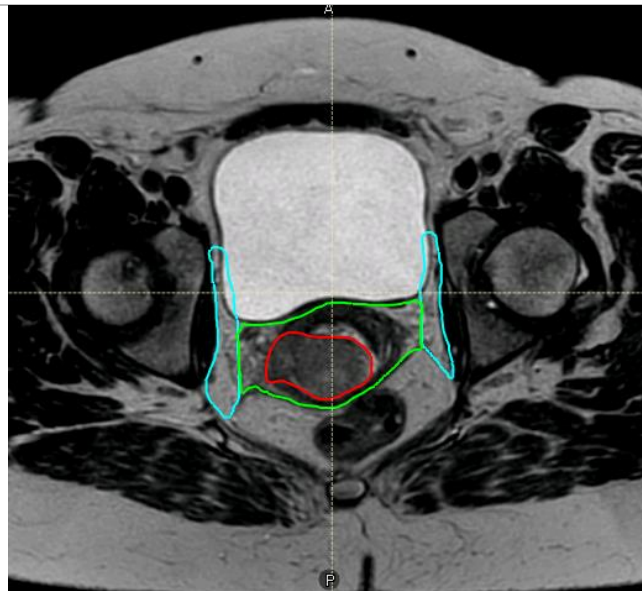
This volume is called the **gross tumour volume** or “**GTV**”; delineated here in red.



Areas of potential microscopic tumour spread are identified. This is termed **clinical target volume** or “**CTV**”.

Expansion around the GTV can be: “*geometric*” – a standard e.g. 10mm expansion, and/or “*anatomic*” – expanding into the anatomical sub-structures with little barrier to spread and therefore most likely to be infiltrated. Most commonly a fusion of these approaches is used.

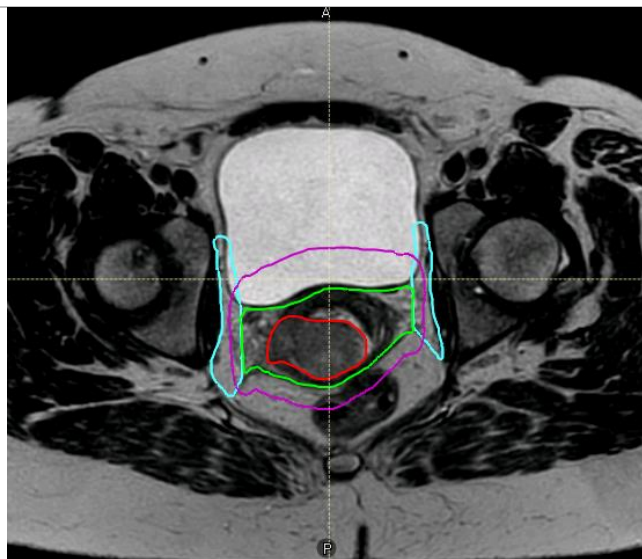
CTV delineated in green (local spread) and blue (lymph node spread)



In some cases, a margin is added to allow for internal movement of the target structures. This is termed the **internal target volume** or “**ITV**”.

Until recently, this has not been commonly used in clinical practice.

ITV (for local target only) delineated in purple



Finally, a further margin is added to account for errors in patient positioning. This is termed the **planning target volume** or “**PTV**”.

This is the volume to which the medical physicists creating the dose plan will shape the radiation dose, and to which the radiation dose will be prescribed.

PTV - the final treatment volume - delineated in blue
(expanded from combined lymph node and local target in orange)

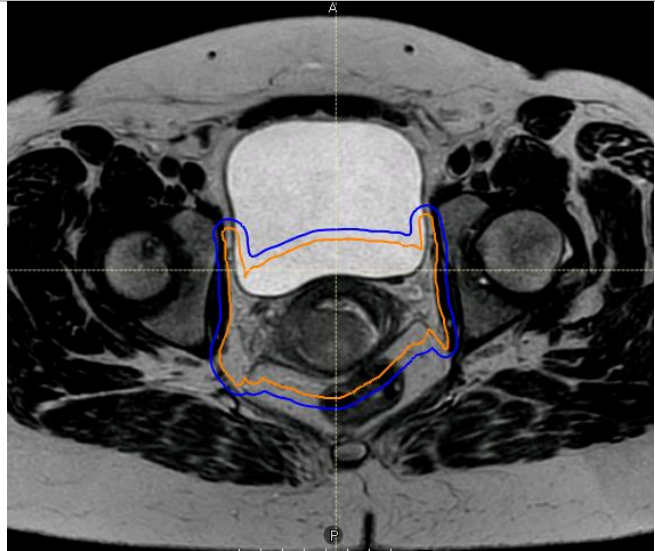


Figure 1-3 - The process of radiotherapy "contouring" or "target volume delineation". Illustrated with contouring of a cervical cancer case for external beam radiotherapy

Target volume delineation forms the basis for the subsequent steps of radiotherapy planning and dose delivery. Crucially, if target volume delineation is incorrect then a systematic error will be introduced for which it is not possible to compensate.

1.3 Selected advances in radiotherapy

Major innovations in the last 20 years of radiotherapy treatment include:

- “Intensity modulated radiotherapy” (IMRT)
- “Image-guided radiotherapy” (IGRT)
- “Image-guided, adaptive brachytherapy” (IGABT)

These innovative radiotherapy treatment techniques seek to enable clinicians to increase the chance of cure by escalating the dose given to the tumour, and/or decrease the dose received by the organs at risk. They are briefly outlined in the following sections.

1.3.1 Intensity Modulated Radiotherapy (IMRT)

Conventional or three-dimensional (3-D) “conformal” radiotherapy involves beams of radiation focused on the tumour from different angles around the body. This results in a radiotherapy dose distribution that is cuboid or polygon-shaped (Figure 1-4).

Intensity modulated radiotherapy allows the radiation dose to be shaped by dynamic motion of the linear accelerator collimator leaves whilst the radiation beam is on. This results in a radiotherapy dose distribution that is more tightly conformed to the tumour:

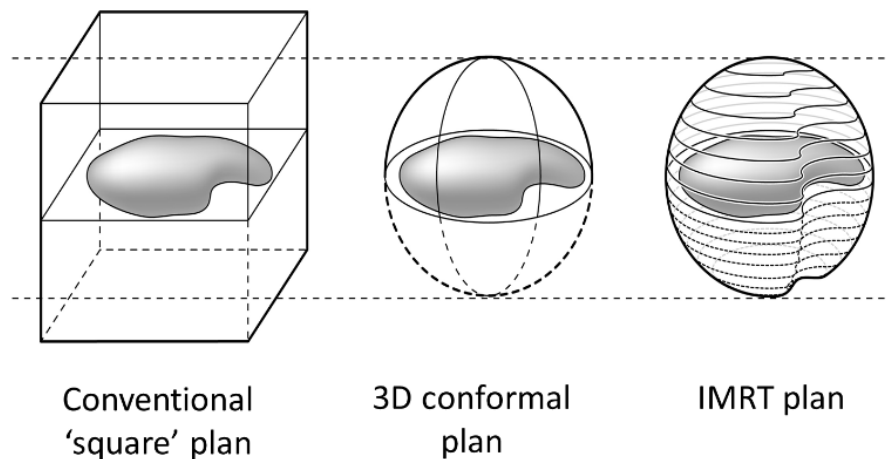


Figure 1-4 - Using advanced radiotherapy techniques to shape the radiation dose. Image courtesy of Dr Raj Jena and Prof. Neil Burnett.

This more precise shaping of the radiation allows a reduction of the dose to the organs at risk. For example in the landmark PARSPORT randomised controlled trial in head and neck cancer, IMRT was used to reduce the dose to the parotid salivary glands. This resulted in a significantly reduced rate of chronic xerostomia (dry mouth) of 29% in the IMRT group compared with 83% in the 3-D conformal radiotherapy group (Nutting et al., 2011).

In cervix cancer radiotherapy, IMRT leads to a reduced volume of irradiated bowel, which is associated with reduced physician- and patient-reported bowel toxicity (Jensen et al., 2021). Preliminary reports from a randomised trial of IMRT versus conformal radiotherapy in gynaecological cancers indicate lower acute bowel toxicity for IMRT (Yeung et al., 2020).

1.3.2 Image-Guided Radiotherapy (IGRT)

Checking the patient is correctly positioned prior to delivery of radiotherapy has always been part of routine clinical practice. Standard practice is to use small tattoos – surface markings that can be lined up with lasers to help reproduce a consistent patient position.

However internal anatomy can move independently of surface anatomy (Roeske et al., 1995). Image-guided radiotherapy is the process of systemically checking this motion and adjusting radiotherapy targeting to ensure that the radiation dose is delivered to the planned target. In a retrospective study of two prostate cancer cohorts (2008-9 and 2006-7), high risk patients

treated with IGRT had increased biochemical tumour control and a lower rate of late urinary toxicity compared with patients treated without IGRT (Zelevsky et al., 2012).

Similar techniques are vital in cervix cancer treatment, where the day-to-day changes in pelvic anatomy (particularly bladder and bowel filling) can significantly affect the position of the target and organ-at-risk volumes (Webster et al., 2020). A systematic review of organ motion for cervix cancer highlighted that the target could move between 5 and 40 millimetres depending on the patient (Jadon et al., 2014). Image-guided methods to compensate for this include patient-specific internal target volume (ITV) or planning target volume (PTV) margins, a 'plan of the day', adaptation of the plan depending on motion during the initial treatment, and live motion tracking (Webster et al., 2020).

1.3.3 Image-guided adaptive brachytherapy (IGABT) for cervix cancer

Radiotherapy with concomitant chemotherapy followed by image-guided brachytherapy is the standard of care for locally advanced cervix cancer (Marth et al., 2017, Cibula et al., 2018) - outcomes are superior to surgery. The first part of the radiotherapy dose is delivered with external beam radiotherapy, usually in 25 fractionsⁱⁱⁱ over 5 weeks with concomitant platinum-based chemotherapy. This induces tumour shrinkage.

The second part of the radiotherapy dose is delivered to the residual disease using brachytherapy in fewer fractions (often 3 or 4) of a higher dose. The advantage of brachytherapy, especially in relatively small tumour volumes such as those after initial treatment in cervix cancer, stems from the inverse square law, which means that tissue closest to the source (inserted into or next to the tumour) receives a very high dose of radiation, which falls off rapidly as the distance from the source increases:

ⁱⁱⁱ A 'fraction' in radiotherapy consists of a single radiotherapy treatment

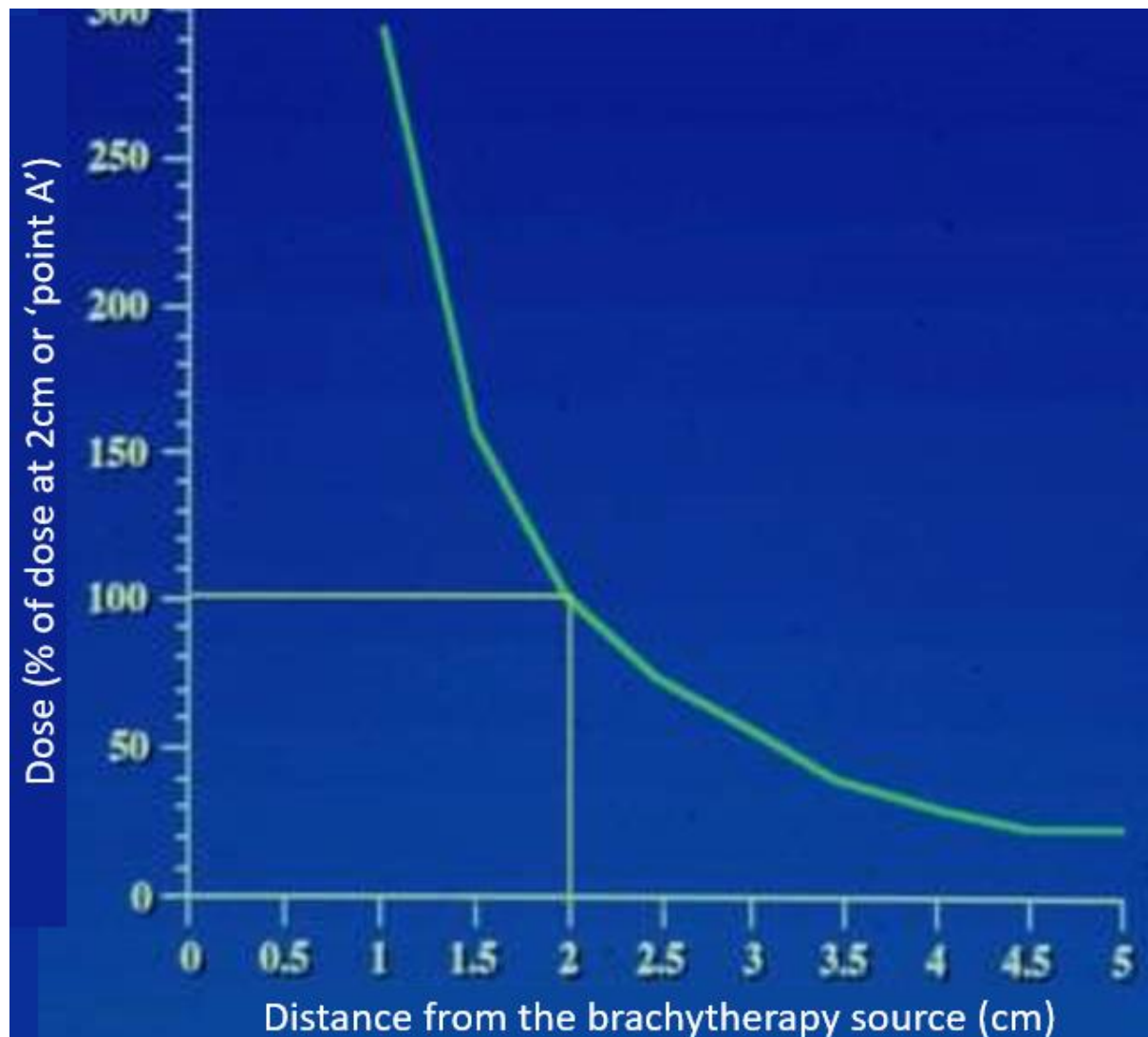


Figure 1-5 - Brachytherapy dose drops sharply as the distance from the source increases. Image courtesy of Dr Li-Tee Tan. 'Point A' denotes the location that brachytherapy dose was prescribed to prior to the era of image-guided adaptive brachytherapy.

This dose distribution means that using brachytherapy with the radioactive source inside the tumour, clinicians are able to give higher doses to the tumour whilst keeping within the radiation tolerance of the nearby organs at risk.

In the 1990s, cervix cancer brachytherapy delivery was based on two-dimensional X-ray imaging using standard approaches developed in the early 20th century. In 2000, a European Society of Therapeutic Radiation Oncology (ESTRO) working group (the "GEC-ESTRO GYN group") was established to support and shape the emerging field of MRI-based adaptive brachytherapy in cervix cancer based on initial experience from a few pioneering European centres (Pötter et al., 2018). The terms 'image-guided' and 'adaptive' refer to the fact that the high soft tissue resolution of magnetic resonance imaging (MRI) at the time of brachytherapy allows 'adaptation' of radiotherapy dose to the *residual* tumour:

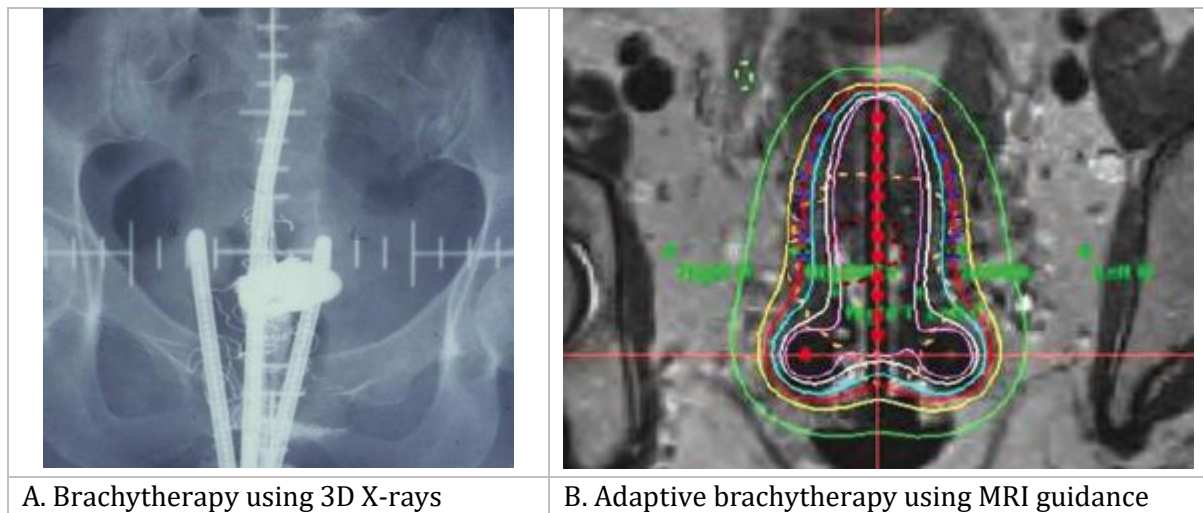


Figure 1-6 - X-ray-guided brachytherapy compared to adaptive MRI-guided brachytherapy. MRI image reproduced from ICRU 89 report, 2013

Out of the GEC-ESTRO GYN collaboration two studies were launched initially:

- Retro-EMBRACE – a retrospective analysis of over 800 cervix cancer patients treated with image-guided adaptive brachytherapy (IGABT) before 2008. This demonstrated high rates of local and pelvic control associated with an overall survival benefit of 10% at 5 years compared to similar historical cohorts treated with standard brachytherapy (Sturdza et al., 2016).
- EMBRACE-I – a prospective observational study of EBRT and MRI-based IGABT. The study closed in 2015 after accrual of 1416 patients. Initial analysis has confirmed that there is a relationship between radiation dose and local tumour control using this advanced radiotherapy technique. It has also demonstrated excellent rates of local and pelvic control and relatively low rates of severe side effects (Pötter et al., 2020).

The EMBACE-II study aims to systematically apply new external beam radiotherapy techniques such as IMRT with daily image-guidance in combination with IGABT in a prospective multi-centre setting (Pötter et al., 2016a, Pötter et al., 2018). The study will be explained further in Chapter 4, as data from the EMBRACE-II trial group makes up a substantial portion of this thesis.

1.4 Radiotherapy contouring – a post-graduate educational challenge

These developments have been adopted into clinical practice relatively recently. Radiotherapy contouring requires skills that were not previously required of oncologists including detailed

interpretation of radiological anatomy (including advanced modalities such as MRI and PET-CT) and predictions of organ motion. However, these skills are not often formally taught in clinical oncology training programmes. Contouring was not specifically mentioned in the 2012 European Society of Radiation Oncology (ESTRO) core curriculum (Eriksen et al., 2012).

During the transition from two-dimensional to three-dimensional ‘conformal’ radiotherapy, conformal radiotherapy (the advanced technology at that time) was sometimes seen to result in a *decreased* rate of tumour control initially, where inadequate imaging and safety margins were used (Kim et al., 1995). A similar learning curve has been seen during the transition from three-dimensional conformal to intensity-modulated radiotherapy – several papers have been published in which early proponents of the advanced technology report recurrences at the edge of the high dose radiation volume (Chen et al., 2011, Eisbruch et al., 2004, Schoenfeld et al., 2008), indicating a geographic miss. Therefore, it is vital to supplement technical innovation with effective education to reduce the learning curve and ensure optimal outcomes for patients.

Consensus guidelines for target volume delineation in many tumour sites have been produced by various professional bodies (Offersen et al., 2015) and trial groups (Michalski et al., 2010), but they are not always consistently interpreted (Ciardo et al., 2017) - inter-clinician variation in contouring has been documented in nearly all tumour sites (see Chapter 2). Given the amount of variation between experienced professionals, setting the standard which learners aspire to, or are assessed against, is difficult.

Currently the main form of summative assessment for radiotherapy contouring available to oncologists is through the radiotherapy quality assurance process in clinical trials, but this is not available to all clinicians. For trainees, contouring is assessed as part of workplace-based assessments. The Royal College of Radiologists^{iv} has plans to introduce a distinct contouring component to their final FRCR examination for trainees (Gwynne et al., 2017).

These issues are discussed further in the literature review in Chapter 2.

^{iv} The Royal College of Radiologists design and oversee training and assessment of Clinical Oncologists in the UK

1.5 Summary

This introduction has outlined the importance of radiotherapy in curative cancer treatment. Recent developments in radiotherapy (IMRT, IGRT, IGABT) have increased cure rates and/or reduced side effects by harnessing advanced technologies.

The process of delivering radiotherapy is complex, with multiple sources of potential error. Radiotherapy contouring is a critical step in this process, as it forms the basis for subsequent design and delivery of radiotherapy. With the increasing complexity of radiotherapy treatment, oncologists require new skills in order to deliver high quality radiotherapy – this is an educational challenge that relates to trainees and established practitioners alike.

2 Radiotherapy contouring variation - a critical review of the literature

2.1 Introduction

As outlined in Chapter 1, advanced radiotherapy techniques improve local tumour control and reduce treatment toxicity by delivering higher radiation doses to tumours while sparing adjacent normal tissue. The benefit of these and other high precision techniques is critically dependent on optimal contouring of the tumour and organs at risk by oncologists, as the steep dose gradients and reduced margins leave little margin for error.

This chapter reviews the literature on contouring variability and its impact on dosimetric and clinical outcomes. The current methods for reducing contouring variability and their limitations are discussed.

Research questions

- What is the magnitude of radiotherapy contouring variability, and what are its dosimetric and clinical consequences?
- What do we know about the causes of radiotherapy contouring variation?
- What strategies can reduce contouring variation and what are their effects?

My contribution to this work

This literature review is based on published work:

- Chang ATY, Tan LT, Duke S, Ng W-T. Challenges for Quality Assurance of Target Volume Delineation in Clinical Trials. *Frontiers in Oncology*. 2017;7.

After Dr Chang produced a first draft, I conducted a further literature search and re-drafted the text significantly, which was then significantly revised by Dr Tan. All four authors reviewed and agreed the final manuscript. The text has been adapted and added to for this thesis.

2.2 Methods and materials

2.2.1 Search strategy

For the initial literature search in 2017, the PubMed and Scopus databases were searched using the keywords “radiotherapy” and (“delineation” or “outlining” or “contouring”^v).

Two recent high quality systematic reviews were identified: one concerning radiotherapy contouring variability (Vinod et al., 2016a), and the other focussing on interventions to reduce this variability (Vinod et al., 2016b). The reference lists of these two articles and associated papers were reviewed. Both studies included articles published until the end of 2014, therefore the initial search results have been updated to include articles published from 2015 to 2020. Relevant abstracts and full texts were reviewed, with studies included in the analysis below if they supplemented the findings from Vinod et al.

In 2019 a further systematic review was published: of educational interventions to improve radiotherapy contouring (Cacicedo et al., 2019). Whilst this mostly refers to studies covered by the previous reviews by Vinod et al., its appraisal and critique of the field is discussed separately below.

2.3 Variation in contouring – magnitude, causes and consequences

2.3.1 Magnitude of contouring variability

The delivery of radiotherapy treatment has long been subject to careful measurement and evaluation of the causes and magnitude of systematic and random errors. As a result, evidence-based strategies have been developed and universally adopted which have enabled radiotherapy delivery to approach millimetre precision (Cubillos Mesias et al., 2016).

In contrast, variability in contouring has not been minimised with the same rigour. In 2016, Vinod et al (Vinod et al., 2016a) published a systematic review of publications on uncertainties in

^v Wildcard (*) expressions were used for these three terms

contouring in radiation oncology. They identified 119 papers on contouring variability published between 2000 and 2014 covering the following clinical topics - breast, bladder, prostate, lung, oesophagus, stomach, pancreas, liver, rectum, head and neck, brain, cervix, uterus, lymphoma, sarcoma, palliative radiotherapy and organ at risk (OaR) contouring. A number of studies focussed on specific advanced radiotherapy techniques including image-guided adaptive brachytherapy for cervical cancer, stereotactic ablative body radiotherapy for lung cancer and stereotactic radiosurgery for brain metastases.

All the studies showed considerable contouring variability between observers, often measured in centimetres. Contouring variability was evident in all the volumes pertaining to radiotherapy planning i.e. the gross tumour volume (GTV), clinical target volume (CTV) and planning target volume (PTV).

Contouring variability is seen amongst experienced radiation oncologists as well as trainees. In one highly cited French study of GTV delineation in lung cancer (Van de Steene et al., 2002), nine radiologists and eight radiation oncologists from five centres were asked to delineate the primary tumour and involved lymph nodes on the computed tomography (CT) images of ten patients. The study reported inter-observer variation in the dimensions of the primary tumour of up to 4.2 cm (transverse), 7.9 cm (cranio-caudal) and 5.4 cm (antero-posterior). The variation in the extreme extensions of the GTV (tumour and lymph nodes) ranged from **2.8 – 7.3 centimetres**. The study showed that compared to radiation oncologists, radiologists tended to delineate smaller volumes and report fewer difficulties for “difficult” cases. Junior doctors delineated smaller and more homogeneous volumes than their senior colleagues, regardless of their specialty, especially for “difficult” cases.

The authors suggested four possible causes for the large inter-observer variation - problems with methodology including definitions and concepts, difficulty differentiating between tumour and benign pathology (e.g. lung collapse), difficulty differentiating between tumour and normal structures, and lack of knowledge of anatomy. Interestingly, they also concluded that only the minority of the issues could be resolved objectively.

2.3.2 Causes of contouring variability

Despite the numerous papers on contouring variability within and outside clinical trials, few have attempted to evaluate the causes of contouring variability in a systematic fashion.

Several studies have reported the impact of imaging modality on contouring variability. For example, a number of studies (Van de Steene et al., 2002, Caldwell et al., 2001, Morarji et al., 2012) showed that more consistent definition of the GTV in lung cancer can be obtained if the CT images were co-registered with 2-[18F]-fluoro-2-deoxy-d-glucose (FDG) positron emission tomography (PET) images. Similarly, there are studies showing more consistent definition of GTV and CTV of brain tumours on CT images co-registered with magnetic resonance images (MRI) (Cattaneo et al., 2005). Image co-registration is now standard practice for both these tumour sites.

Reduced contouring variability seen on one imaging modality does not necessarily equate to this being a superior imaging modality. In a study on image-guided adaptive brachytherapy for cervical cancer (Viswanathan et al., 2014), 23 gynaecologic radiation oncology experts were asked to delineate the CTV on CT and MRI. There was a higher level of agreement of contours on CT despite MRI being universally recognised as the superior imaging modality. This probably reflects clinician unfamiliarity with MRI image interpretation for IGBT cervix planning, where post-radiation changes can be a confounding factor.

It is commonly assumed that the major cause of intra-observer contouring variability is suboptimal image interpretation (Riegel et al., 2006). However, other factors such as conceptual understanding of patterns of tumour spread and organ motion are equally important. In a study on definitive radiotherapy for cervical carcinoma (Weiss et al., 2003), five radiation oncologists and two gynaecologists independently contoured the CTVs for three patients. The study showed good consistency in outlined anatomical structures suggesting that image interpretation was not an issue. However, there was large inter-observer variability in CTV delineation: *“the ratio between largest and smallest volumes ranged between 3.6 and 4.9 for all observers ... The ratio of common volumes to encompassing volumes ranged between 0.11 and 0.13 for the radiation oncologists, and between 0.30 and 0.57 for the gynaecologists”* (Weiss et al., 2003, p.87).

The contouring variability between gynaecologists and radiation oncologists probably reflects different conceptual understanding of areas at risk of microscopic disease between the two specialties. The core skill for gynaecologists is to remove the tumour with a small margin (usually 5 mm) with minimal disruption of surrounding tissue. In contrast, radiation oncologists irradiate large volumes of tissue to a relatively homogenous dose to minimise the risk of in-field and edge-of-field recurrences. The concepts of microscopic disease for these two specialties are therefore likely to be very different. This explanation could also account for the contouring variability between radiologists and radiation oncologists in the lung cancer study. Cancer radiologists are required to accurately define the tumour (avoiding both under and over estimation) to predict surgical resectability whereas the prime concern of radiation oncologists is to avoid missing the

gross tumour and microscopic spread. It is therefore easy to see why in difficult cases, some radiation oncologists would err on the side of caution and include areas of uncertainty in the GTV.

It is also well recognised that junior doctors are less able to appreciate uncertainties than their senior colleagues, a phenomenon known as the Dunning-Kruger effect (Kruger and Dunning, 1999) based on Charles Darwin's quote that "ignorance more frequently begets confidence than does knowledge".

Consistency and clarity of conceptual understanding is particularly important when new concepts are introduced. An example is the internal target volume. The margin for the ITV (called the internal margin) is distinct from the set-up margin used for the PTV. However, in a survey of 50 radiation oncologists at a pelvic IMRT workshop (personal communication - Dr Li-Tee Tan), 38% did not use the concept of the ITV in their daily practice, 30% incorporated the internal margin into the CTV, 26% incorporated the internal margin into the PTV and only 8% contoured the ITV as a separate structure.

2.3.3 Dosimetric and clinical impact of contouring variability

There are no studies which have assessed the direct impact of contouring variability on clinical outcome. Some studies model the impact of contouring variability on radiation dose ("dosimetric") parameters as a surrogate of clinical impact, as there is evidence in nearly all tumour types that reduced radiation dose leads to a reduction in tumour control.

Dosimetric impact

The Vinod review identified only 25 (21% of the total) studies which evaluated the impact of variability in target and OAR delineation on dosimetry (Vinod et al., 2016a). Thirteen studies evaluated the dosimetric impact of target volume variability; it was interesting that three of these studies found no significant impact on PTV dose coverage.

Van de Steen et al. (Van de Steene et al., 2002) estimated the impact of GTV delineation variability on tumour control probability (TCP) using the dosimetric variation data and modelled dose to tumour control relationships. Across all plans, the mean TCP decreased from 51% for a matched plan (i.e. a plan created for that GTV volume) to 42% for an unmatched plan (i.e. a plan created for another GTV), a difference of 9%. The mean range in TCP across the eight patients was 2% (maximum range 5%) for matched plans compared to 14% (maximum 31%) for unmatched plans. They also estimated the normal tissue complication probabilities for different OAR but this

analysis was of limited value as the plans used were 4-field boxes (i.e. basic 'conformal' radiotherapy) which would not have been used clinically.

Jameson et al (Jameson et al., 2014) also modelled the impact of GTV delineation variability on TCP and equivalent uniform dose in lung cancer. Three radiation oncologists contoured the GTV on the planning CT, the diagnostic PET-CT and the radiotherapy planning PET-CT for seven patients. An optimised plan with 3-5 conformal beams was created for each volume. The standard deviation of the volumes across all seven patients ranged from 39-419 cc. However, the standard deviation of the equivalent uniform dose was ≤ 1 Gray^{vi} (Gy) in 4 of the 7 patients (range 0.09 – 21.2 Gy). Similarly, the standard deviation of the TCP was negligible (0-1%) in 4 of the 7 patients (range 0-22% over all patients). Contouring variations in the lateral dimensions had the greatest impact on TCP.

Clinical impact

There are numerous reports in the literature of suboptimal radiotherapy contouring, which can lead to fatal marginal recurrences due to geographical miss (Chen et al., 2017, Eisbruch et al., 2004, Chen et al., 2011, Schoenfeld et al., 2008).

Wider-scale data come from observational studies of radiotherapy quality assurance. In a landmark study, Peters et al (Peters et al., 2010) retrospectively analysed the radiotherapy plans of 780 patients in the Trans-Tasman Radiation Oncology Group 02.02 (TROG 02.02) HeadSTART trial in head and neck cancer. They found that patients whose radiotherapy plans failed trial quality assurance but were not corrected (12% overall) had poorer survival and loco-regional control compared to the those with protocol-compliant plans (2-year overall survival (OS) 50% vs. 70%, $p < 0.001$, 2-year loco-regional control 54% vs. 78%, $p < 0.001$):

^{vi} Gray - the unit of ionizing radiation dose

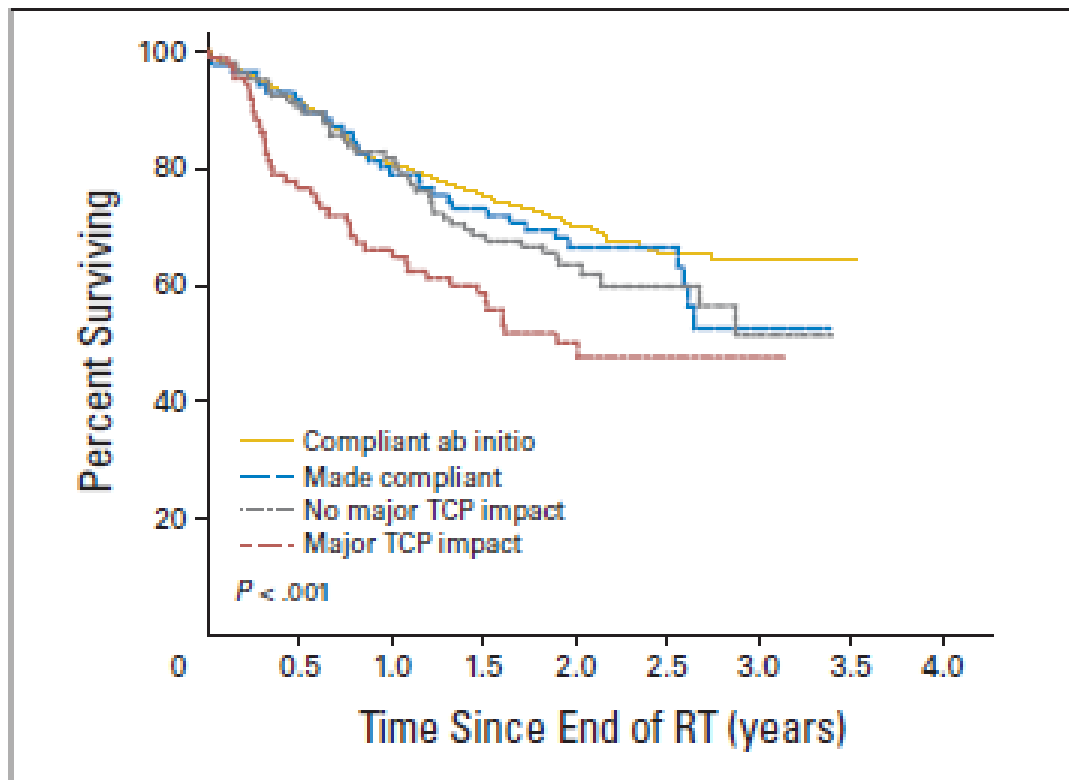


Figure 2-1 - Patient survival in the TROG 02.02 study in head and neck cancer, by compliance to radiotherapy protocol and tumour control probability (TCP) classification

However, incorrect volume delineation was a feature in only 25% (24/97) non-compliant plans. This demonstrates that although contouring is important, it is by no means the only skill in radiotherapy which should be targeted for educational intervention.

A meta-analysis of radiotherapy protocol deviations (Ohri et al., 2013) reported a statistically significant increase in the risk of death associated with radiotherapy protocol deviations (Hazard ratio (HR) of death = 1.74; 95% CI = 1.28 to 2.35, $P < 0.001$), but it is important to remember that correlation is not causation. The meta-analysis included a randomised study of adjuvant radiotherapy for pancreatic cancer (Abrams et al., 2012) which reported a HR of death with radiotherapy protocol deviation of 1.33 after adjustment for prognostic factors, but adjuvant radiotherapy in pancreatic cancer does not improve survival (Neoptolemos et al., 2004, Ducreux et al., 2015). Within the study there was no evidence of worse loco-regional control or fatal toxicity in the protocol deviation cohort, so the cause for the lower survival is unclear. The authors speculated that an unmeasurable aspect of improved loco-regional control led to better survival in the per protocol cohort, but it may be due confounding prognostic factors associated with protocol deviations (such as surgical quality or unadjusted patient factors).

Nevertheless, the overall findings from case series, dosimetric data and clinical trial RTQA provide a compelling picture of the importance of high-quality radiotherapy and the potentially disastrous impact of errors in contouring.

2.4 Minimising contouring variability : strategies

Several interventions have been developed to reduce inter-observer contouring variability. These were reviewed in the second publication by Vinod et al. (Vinod et al., 2016b) .

2.4.1 Contouring guidelines and atlases

The most common method for reducing contouring variability within and outside clinical trials is probably the use of consensus contouring guidelines and/or atlases (Lobefalo et al., 2013, Vinod et al., 2016b, Lin et al., 2020).

Lobefalo et al. evaluated the benefit of a contouring guideline on consistency of contouring in a study of rectal cancer. Four radiation oncologists contoured the CTV on 10 patients before and after the introduction of a shared guideline. The Agreement Index (a measure of geometric overlap) improved from 0.57 (pre-guideline) to 0.69 (post-guideline). The unmatched PTV coverage improved from 93.7 +/- 9.2% to 96.6 +/- 4.9% for 3D conformal radiotherapy and 86.5 +/- 13.8% to 94.5 +/- 7.5% for a volumetric modulated arc radiotherapy (VMAT) technique. This suggests that the dosimetric impact of inter-observer variation is more pronounced for advanced radiotherapy techniques.

Eminowicz et al (Eminowicz et al., 2016a) from the INTERLACE trial reported the reduction of inter-observer contouring variation and increased protocol adherence after introduction of an atlas. They analysed seven key guidelines for target volume contouring in cervical cancer and identified 11 common areas of variation. A pictorial atlas was then derived to illustrate a consistent delineation method for these areas. The average proportion of outlines (of 4: primary CTV, nodal CTV, bladder, rectum) complying to the protocol improved from 1.8/4 to 2.7/4 with atlas use.

Gillespie et al have used interactive, dynamic and responsive web-based technologies to enhance the transmission of information from a radiotherapy atlas (Gillespie et al., 2017). In a multi-centre randomised trial of interventions for radiation oncology trainee contouring, participants were randomised to re-contour a case of nasopharyngeal cancer using “currently available resources” (“including textbooks, review articles, trial protocol descriptions, and consensus guidelines”) or

a novel web-based 3D contouring atlas (see <https://econtour.org>). Those using a web-based atlas showed greater agreement with the consensus and expert contour for the parotid gland (DICE conformity index [a measure of geometric overlap] 0.63 vs 0.52; $p = 0.02$) and reported higher usability scores and satisfaction with the material.

While contouring guidelines are undoubtedly invaluable in making contouring more consistent, they can also be a source of variability if different groups produce conflicting guidelines for the same tumour site or anatomical region. For example, the GYN consortium consensus guidelines for CTV delineation for IMRT for cervix cancer defines the lateral border of the parametrium as the medial edge of internal obturator muscle/ischial ramus (i.e. lateral to the pelvic vessels) whereas the EMBRACE-II guidelines define this border as the medial edge of internal iliac and obturator vessels. Similarly, the inferior border of the pre-sacral nodes have been defined as S2 in gynaecological guidelines (Small et al., 2008), S3 in prostate cancer guidelines (Michalski et al., 2010) and bottom of the coccyx in anal guidelines (Muirhead et al., 2016). One can see how a clinician used to contouring in a particular way will continue to do so in a clinical trial regardless of the protocol specification.

2.4.2 Multi-modality imaging

Improved imaging, e.g. use of intravenous contrast, optimal window settings and multi-modality imaging, is an intuitive way to improve contouring consistency. In the Vinod review (Vinod, Min, et al., 2016), there were more published studies using this method than all other methods combined. However, results have been mixed and 9 of the 31 studies reviewed did not demonstrate a statistically significant reduction in contouring variability. It appears that interpretation of the additional imaging modality and image co-registration are sources of error in themselves.

2.4.3 Auto-contour provision

A few studies have reported improved contouring consistency from clinicians editing an auto-contour compared to manual delineation (Vinod, Min, et al., 2016).

However, if the auto-contour contains an error, then this is more likely to be transmitted through the manual editing process as a systematic error. The majority of auto-contouring software in clinical use as of 2016 utilised atlas-based segmentation which requires manual editing due to the wide variation in normal and post-treatment anatomy. Machine learning techniques hold promise for increasing accuracy and reducing the burden of user editing as discussed in a review

by Sharp et al (Sharp et al., 2014). They are not yet in routine clinical use, though, and still require clinicians to take responsibility for the contours and treatment plan. By definition they cannot innovate to produce new target concepts. Therefore clinicians still need to know how to contour radiotherapy target volumes and organs at risk.

2.4.4 Contouring workshops and educational programmes

Contouring workshops are a popular method for teaching contouring but they have several limitations.

Cacicedo et al. conducted a systematic review of educational interventions to improve radiotherapy contouring. 16 studies representing 370 participants (average 23 per study; range 4-141) were identified. These were studies of contouring simulation training (see Chapter 3, Section 3.4), although they did not always self-identify as such. In most cases improvement in contouring skill was measured by re-contouring on the same case; most commonly (in 7/16 studies) improvement was measured by analysing the geometric similarity of participants' contours to a gold-standard contour. Only one study performed a qualitative evaluation of contours (i.e. assessed clinical adequacy) and only one study evaluated the impact beyond 6 months. The large majority of interventions were "one-off" rather than programmes of learning and assessment. It is therefore difficult to understand the degree of clinically relevant improvement and to ascertain whether learning was "retained" and could be "transferred" (see Chapter 3) to different cases with different patient anatomy and tumour topography.

One example of a longer-term educational intervention is an International Atomic Energy Agency study over a one-year period, involving 11 pairs of clinicians comprising a radiation oncologist and a nuclear medicine physician (Konert et al., 2016). Training consisted of lectures, contouring practice, and group and individualised feedback. After the first training session, geometric overlap indices for three repeated cases increased from $0.57(\pm 0.07)$ to $0.66(\pm 0.07)$. After further training, geometric overlap for the same three cases further increased from $0.64(\pm 0.06)$ to $0.80(\pm 0.05)$, $p = 0.01$.

Recent advances in technology such as web-enabled video conferencing and interactive software have enabled both live and offline educational interventions to reach across geographical boundaries. An example is the FALCON programme (Fellowship in Anatomic delineation and Contouring), offered by the European Society for Radiotherapy & Oncology (ESTRO) (Eriksen et al., 2014). Online workshops, however, will face the same pedagogical issues as live ones.

Several contouring tools have been developed to support self-learning contouring programmes (for an example see <https://proknowsystems.com/quality/contouring>). These tools offer delineation practice often with provision of a reference volume and/or automated quantitative feedback. These programmes are in their infancy and their utility remains to be established. Issues include difficulty in defining a reference volume given the extent of disagreement in contouring among experts, challenges for user engagement and outdated internet access - particularly in hospitals.

2.4.5 Peer review

Peer review involves the review of aspects of radiotherapy treatment by two or more radiation oncologists, or another specialist such as a radiologist. It may cover indications for treatment, treatment approach, volume delineation, planning directives, evaluation of plan quality and/or treatment verification. The American Society for Radiation Oncology (ASTRO) has identified contouring as the first priority for peer review due to the heterogeneity in contouring and its impact on the rest of the radiotherapy process (Marks et al., 2013).

Multiple audits of peer review have identified that a proportion of radiotherapy treatments require significant alteration. In an early study (Brundage et al., 1999), 3052 cases were reviewed over 8 years of which 4.1% were not clinically acceptable. More recently Mackenzie et al. (Mackenzie, Graham, & Olivotto, 2016) presented a prospective audit of peer review meetings in breast, head & neck and lung cancer. Overall 9% of treatments required alteration before the first or next fraction of radiotherapy, although this varied significantly across the tumour sites (1 – 16%). A study by Dimigen et al. (Dimigen, Vinod, & Lim, 2014) reported that involving a radiologist in weekly quality assurance meetings resulted in a significant change in management in 6% of cases.

Multiple professional organisations now advocate peer review as an important component of safe and effective radiotherapy. However, there are significant barriers to its implementation including a lack of personnel, dedicated time and facilities, and a reluctance of clinicians to invite scrutiny, especially across institutions. In addition, peer review does not guarantee an improvement in contouring. In a study of contouring peer review in lung cancer by 22 trainees, Mercieca et al. noted that major improvements were made by six participants during peer review, but one erroneously excluded part of the tumour and another missed a major error (Mercieca et al., 2020). Rigorous training is vital to avoid a situation where the 'blind are leading the blind'.

Given its cost and resource implications, rigorous research to evaluate its benefit is urgently needed. Technologies which allow large-scale remote assessment of contours would be advantageous.

2.4.6 Minimising contouring variability in clinical trials

The process for radiotherapy quality assurance (RTQA) of contouring in clinical trials may involve one or more of the following:

- **A benchmark case** - the participating institution is asked to delineate radiotherapy volumes on one or more standardised cases according to the protocol.
- **A dummy run** - the institution uploads the datasets of one or more of their previously treated patients for central review.
- **Individual case review** - during the course of the trial, some or all of enrolled patients' radiotherapy datasets are reviewed centrally. This can be prospective or retrospective.

Many reports on RTQA for contouring have used benchmark cases. The general structure of accreditation to a trial using benchmark cases is shown below:

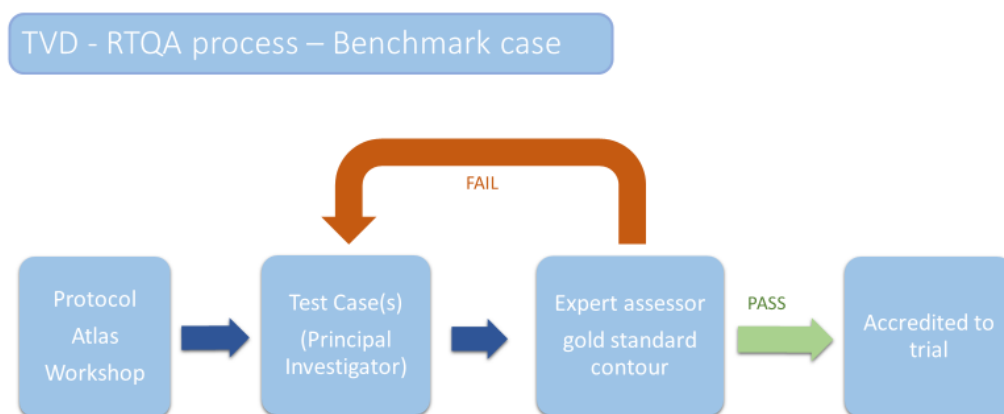


Figure 2-2 - Flow diagram of the radiotherapy quality assurance (RTQA) process for a benchmark case

One example is the INTERLACE study on IMRT for cervix cancer (Eminowicz & McCormack, 2015). Twenty two principal investigators (PIs) from participating centres were asked to contour the CTV on two cases with different FIGO stages. The delineated volumes ranged from 340 cc to 676 cc for case 1 and 458 cc to 806 cc for case 2. The direction of the maximum variation was different in the two cases.

The EMBRACE-I study on IGBT for cervix cancer is an example of RTQA based on a dummy run (Kirisits et al., 2015). Each centre was asked to upload a “good response” case and a “poor response” case for central review. The review was qualitative and quantitative with one physician reviewing all the external beam radiotherapy (EBRT) contours and three other physicians reviewing the brachytherapy (IGABT) contours. Out of 30 submitting centres, 13 had major inconsistencies in BT contouring while 11 had major inconsistencies in EBRT contouring. Centres with experience in IGABT (>30 cases) performed better than those with limited experience.

Retrospective individual case review was reported by the SCALOP trial in pancreatic cancer (Fokas et al., 2016). The chief investigator and a radiologist contoured the GTV on the 60 of 74 patients who received radiotherapy in the study (12 patients had planning CTs which were deemed to be of insufficient quality for re-contouring) and compared their gold standard contours with the treating clinicians’ contours using the Jaccard conformity index and geographical miss index. The agreement (by conformity indices) between expert and clinician contours was better for live contours than for the pre-trial benchmark case, suggesting that the RTQA process may have improved contouring skill. However, errors were still seen - the tumour was completely missed in 1 case, and $\geq 50\%$ of the tumour was missed in 3 cases. Patients for whom the overlap between clinicians’ delineation (Jaccard conformity index) for GTV was greater than 0.7 were significantly more likely to experience cancer progression, a result which the authors describe as counter-intuitive. In this case, there may have been greater agreement on larger tumours which are potentially easier to delineate. The relationship between geometric conformity and contour quality is therefore not straightforward.

2.5 Discussion

This review has found that although there are numerous publications reporting considerable contouring variability within and outside clinical trials, there are very few which have investigated the underlying causes of the variability or its impact on actual clinical outcomes. The limited data on outcomes are conflicting with modelling papers suggesting different impact on tumour control probability in different patterns which is perhaps not surprising. All the data to date suggest that the relationship between contouring variability and outcome is not straightforward and further research is required.

Similarly, several educational strategies have been put forward to minimise contouring variability but there is little systematic research into the effectiveness of the strategies, whether

learning is retained and which underlying constructs they are attempting to engage with and/or modify.

The logistics of radiotherapy clinical trials are such that most trials limit their RTQA process to the principal investigators (PIs) who are probably the most likely to contour correctly. Similarly, most RTQA is based on 1 or 2 carefully chosen benchmark cases which does not take into account the variation in patient anatomy and tumour topography seen over the breadth of the clinical trial population.

The assumption that once the PI passes the assessment, all target volumes on future cases will conform to the standards of the radiotherapy protocol is dubious, and is contradicted by a recent analysis of live case reviews in the neo-SCOPE trial (Evans et al., 2018), showing that cases delineated by clinicians who have passed RTQA contain errors. In large clinical trials timely (or even retrospective) review of all patient radiotherapy treatments, even if desired, is simply not feasible due to resource constraints. There may also be a conflict of interest for the central review team to pass centres in order to increase trial recruitment. In undergraduate final examinations, it would be unthinkable to re-test failing candidates on the same cases, but the resources required to arrange re-testing in examinations are considerable (Pell et al., 2013). These problems are magnified in clinical practice where resources for quality assurance of contouring are generally even more limited.

Given the degree of variability of target volume delineation, even amongst experts, it is unsurprising that there is no consensus on how to systemically assess target volume delineation. Participants are often assessed against a consensus contour or a contour derived via the STAPLE algorithm (Warfield et al., 2004) given the difficulty of estimating the 'ground truth' contour for any given case. There have been no documented assessment criteria in radiotherapy contouring assessment, with most papers simply reporting the percentage of participant contours passing or failing, and there is often no formal consensus as to what represents significant versus non-significant discrepancy.

Those instigating clinical trials in radiotherapy are likely to be domain-specific experts but may have limited experience in crafting robust assessments. Given the progress made in assessment in undergraduate and mainstream postgraduate education in recent decades, recommendations outlining how best practices in assessment pertain to radiotherapy contouring quality assurance would be helpful. Creating these would require a diversity of expertise including: postgraduate assessment, a spectrum of clinical subsites, and radiotherapy quality assurance. Organisations with a mission to standardise radiotherapy, such as the Global Quality Assurance of Radiation

Therapy Clinical Trials Harmonization Group (Melidis et al., 2014), would be well suited to such a task. The group's publications^{vii} show a focus on technical factors as opposed to human factors, which is not surprising given radiotherapy's foundation in the physical sciences, however the importance of human factors in radiotherapy quality and safety is increasingly recognised (Chan et al., 2010, Huq et al., 2016). A systematic review of reported assessment methodology would be helpful in assessing past (and current) practices and may provide a platform for recommendations to improve the process.

2.6 Summary

A large number of studies covering many different tumour sites have documented significant inter-observer variation in radiotherapy target volume delineation. There is indirect evidence of a moderate impact on long term tumour control from radiotherapy dose-modelling studies and retrospective analysis of radiotherapy quality within clinical trials.

Several types of intervention have been trialled in order to reduce this variation – of these, radiotherapy atlases and workshop-based interventions have most consistently shown a reduction in contouring variation. Very few studies have evaluated medium or long-term impact.

Summative assessment of contouring has largely taken place within RTQA thus far, with no documentation of formal assessment criteria. The scope of contouring assessment is limited by resources. Development of an automated formative (or even summative) assessment of radiotherapy contouring would allow testing of all trial clinicians over a more representative range of clinical scenarios, strengthening the validity of the RTQA accreditation process. An online educational programme utilising this kind of formative assessment would have the potential to improve standards and thereby patient outcomes on a large scale.

^{vii} <https://rtqaharmonization.org/publications/>

3 Identifying and applying relevant educational theory

3.1 Introduction

The issues highlighted in Chapter 2 such as assessment of competency, optimal instructional design, and effective simulation have been wrestled with for many years by educationalists both within and outside medicine. Before probing further within the radiotherapy field, in order to avoid “re-inventing the wheel”, it is important to ask the questions:

- **“How can medical education literature and the wider educational literature regarding simulated practical skills training inform our approach to the teaching and assessment of radiotherapy contouring?”**, and
- **“how can this knowledge be applied to shape simulated assessment and teaching of radiotherapy contouring?”**

Comprehensive systematic search strategies are prized even in complex interventions (Craig et al., 2008) due to their objective search criteria and reproducible methods, but may not be appropriate for every research question (Greenhalgh et al., 2018). The questions above are addressed by vast fields of literature - tens to hundreds of studies per year report results of simulation interventions in medical education alone (Cook, 2014). Scoping reviews are a potential alternative to the burden of a full systematic review (Arksey and O'Malley, 2005, Bing-You et al., 2017) but even a scoping review of such broad topics would not be feasible within the time & resource limitations of a doctoral research programme.

In addition, systematic methods are not completely free from bias (Eva, 2008, Greenhalgh et al., 2018) and can be unsuited to addressing questions of what literature is relevant to doctoral research (Maxwell, 2006). In limiting the search to selected populations (e.g. medical doctors versus other healthcare professionals versus non-healthcare; graduate versus undergraduate learners) interventions, comparisons, or outcomes as suggested in systematic review guidelines (Higgins et al., 2019), researchers may miss relevant approaches or perspectives, and fail to see the spectrum of possible interventions within their broader contexts (Greenhalgh et al., 2017).

In other words a focussed literature review, however thoroughly conducted, can involve a form of bias (Chandra et al., 2008): *“It is more much more valuable if the researcher considers the literature broadly in order to fundamentally redefine the way the focal question is conceived in a*

meaningful and insightful manner, rather than going to elaborate lengths to establish that every paper relevant to a very narrow question has been considered" (Eva, 2008, p.853). Similarly, in advocating for the use of non-systematic reviews when appropriate for the research question Greenhalgh et al. *".. distinguish between problems that require data (for which a conventional systematic review, with meta-analysis where appropriate, may be the preferred methodology) and those that require clarification and insight (for which a more interpretive and discursive synthesis of existing literature is needed)"* (Greenhalgh et al., 2018, p.2). I propose that the latter form of knowledge synthesis is required in this instance.

3.2 Methods and materials

I sought to address my research questions by initially reading the educational literature broadly but superficially, and then choosing relevant topics to explore in greater depth. The **hermeneutic process** is suited to these requirements, where *"reading, conducting empirical research and writing are not a linear but rather an iterative process"* and involve *"making sense of a potentially large body of literature relevant for a targeted problem"*(Boell and Cecez-Kecmanovic, 2014, p.259-60).

The hermeneutic approach originated in the 19th century and was concerned with reconstructing the original meaning of biblical texts in their context (Schleiermacher, 1838) but was then expanded more widely to all textual interpretation and subsequently to general understanding (Heidegger, 1929) in domains such as law and medicine (George, 2020). Drawing on this approach, Boell and Cecez-Kemanovic have constructed a hermeneutic framework for the literature review involving two cycles: an initial phase of search and acquisition followed by subsequent analysis and interpretation which then leads to further searching (Figure 3-1):

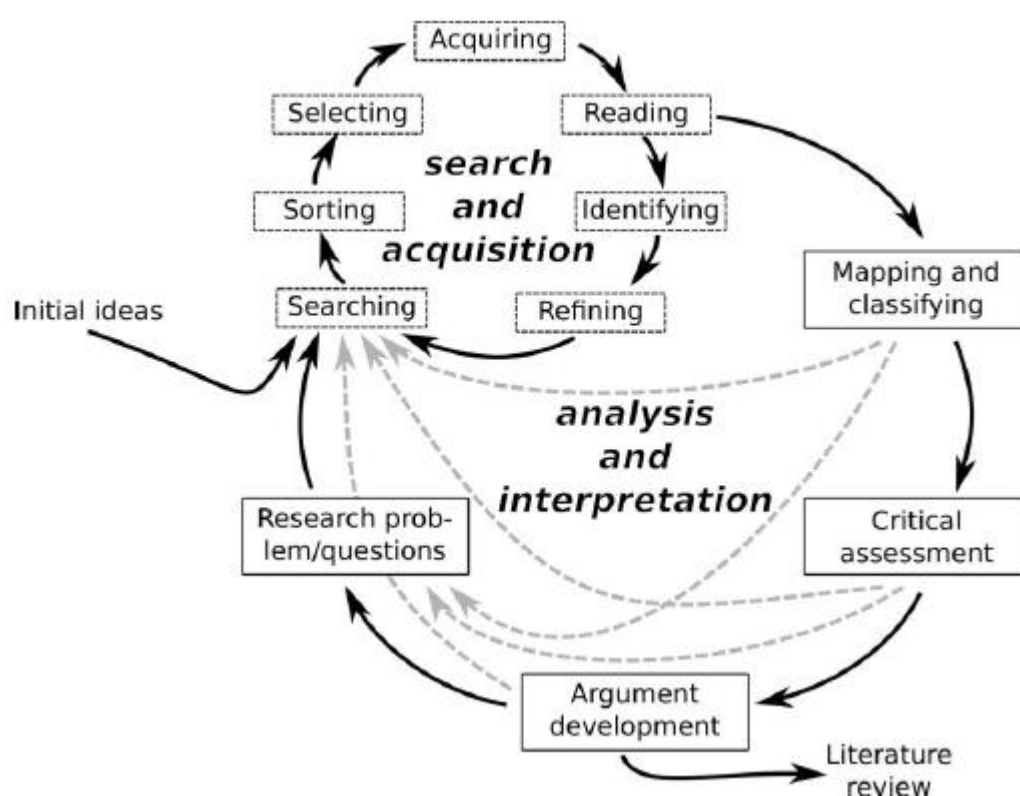


Figure 3-1 - Hermeneutic framework for literature reviews. Reproduced from Boell & Cecez-Kemanovic (2014), p. 264

In the search for breadth of scope and understanding, this approach comes with obvious limitations, especially regarding researcher bias. This method is not inherently reproducible (or least anywhere near the extent of a systematic review), and can never be truly 'completed' - so the decision of when to interrupt the cycle and move on is a pragmatic decision unique to the researcher's goals (Boell and Cecez-Kecmanovic, 2014). To counter these one can: reflect on one's own bias (Cohen et al., 2017); seek multiple perspectives (particularly for critique) from supervisors, networks and colleagues (Greenhalgh et al., 2018); and document decisions about the limitations of depth & breadth which can subsequently be re-examined (Whitehead, 2004).

There is no comprehensive taxonomy of medical education terms to guide the type of search envisaged, although previous attempts have been made (Haig et al., 2004, Haig et al., 2005). The search was focussed on **developing practical skills in graduate medical education with a focus on simulation**, but sought to incorporate relevant research from the undergraduate medical and wider non-medical educational literature where it supplemented or complemented findings from medical education.

The search commenced in 2017 with a review of medical education textbooks, which can *“provide a wider perspective on the research topic”* than journal publications alone (Boell and Cecez-Kecmanovic, 2014, p.278); these textbooks are listed in Appendix Table A.3-1. Textbook chapters were initially skimmed and relevant concepts identified, with those topics being read in full. The second iteration of the literature search used Pubmed, Scopus and Google Scholar to identify reviews (both systematic and non-systematic) of the concepts deemed to be most relevant to the research question. At this stage, concepts were sorted into priority, intermediate and low relevance. Snowballing and selected citation tracking (Greenhalgh and Peacock, 2005) was then employed to identify relevant original research and reviews. Regular review of the 4 highest impact medical education research journals (Appendix Table A.3-2) was conducted after 2017.

In order to be awarded a doctoral degree, students must demonstrate *“the creation and interpretation of new knowledge, through original research or other advanced scholarship, of a quality to ... extend the forefront of the discipline”* (The Quality Assurance Agency, 2014). The original scholarship performed here is not in the critique of the relevant bodies of literature but in their identification, collation, and application to radiotherapy contouring.

3.3 Results

Figure 3-2 shows a thematic map of the domains found in the initial search through medical education textbooks with these divided - according to my own judgement - into lower and higher relevance for simulation of practical clinical skills.

Under the umbrella of simulation, I chose four further domains to explore in greater depth in the first ‘analysis and interpretation’ hermeneutic cycle:

- **Cognitive load theory**
- **Deliberate practice theory**
- **Assessment**
- **Feedback**

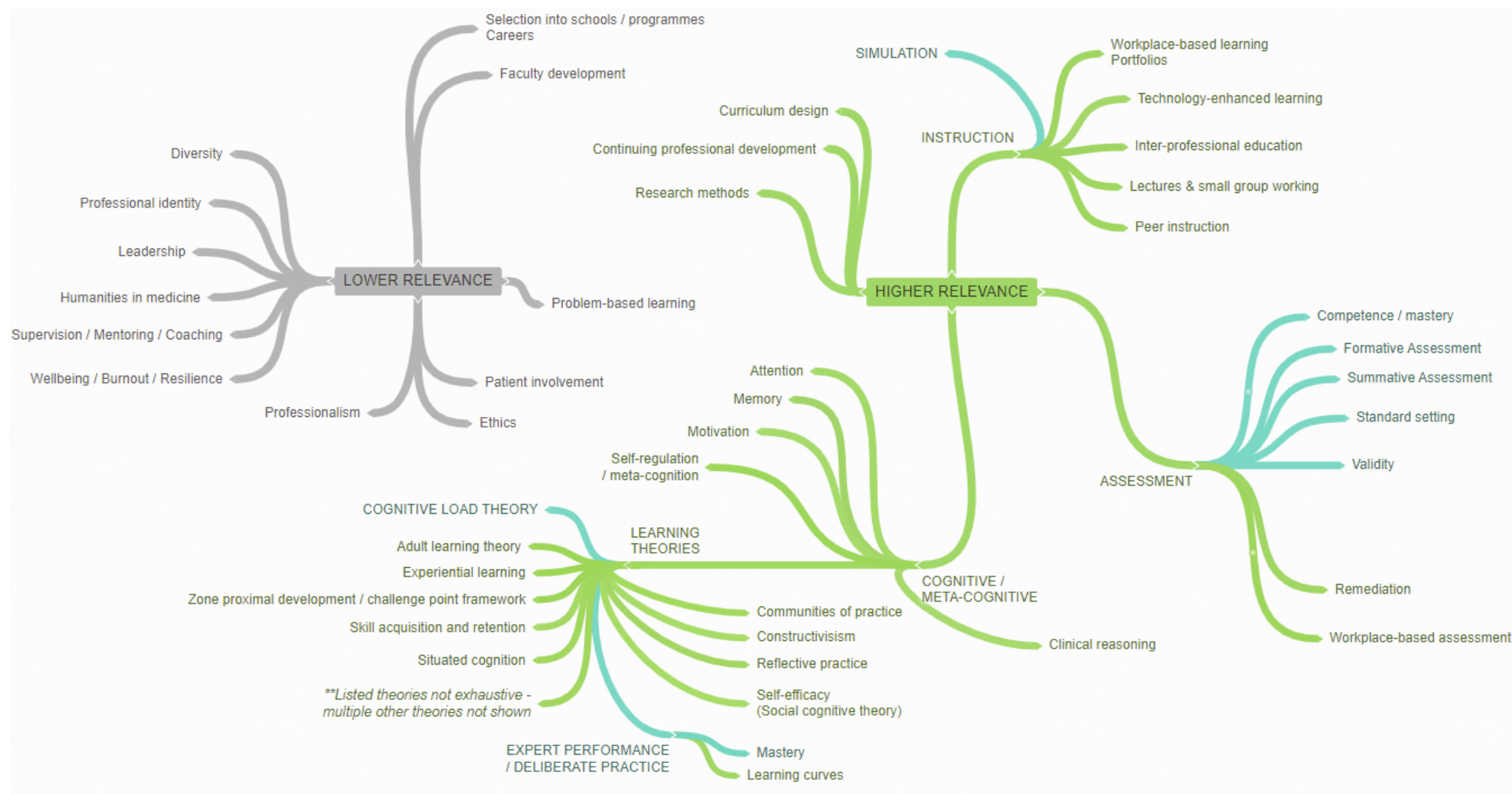


Figure 3-2 - Thematic map of domains identified from a search of medical education textbooks, arranged by relevance to simulated assessment and teaching of practical skills. The domains selected for deeper exploration in the first hermeneutic cycle are shown in turquoise.

I chose these domains due to the overlap between them when considering learners, simulation technology, simulated assessment, and simulated learning exercises. In the first hermeneutic cycle I excluded instructional modalities other than simulation, the wider learning environment and/or programme (e.g. curriculum, workplace, and group or interprofessional learning), specific theories of skill acquisition, clinical reasoning, and meta-cognitive aspects (e.g. motivation and self-regulation). This is not to underestimate their importance in simulated learning and assessment, but to say that they can be explored in greater depth at a later stage when it is clear in what context a simulated assessment or teaching programme is going to be placed.

As discussed in the methods section inevitably this was a subjective judgement, but it is one that could not be avoided given the balance between the scope of in-depth analysis and the available resources.

In the remainder of this chapter, I begin with simulation as the overarching domain (with a focus on practical skills), and then explore the four other themes. Each section will explore the topics in order identifying their:

- definitions and context,
- validation in health professions education and/or best practices,
- areas of uncertainty, critique, or limitations, and
- potential applications to radiotherapy contouring education.

3.4 Simulation

3.4.1 Definition & context

Simulation-based learning is defined in the Dictionary of Simulation in Healthcare as:

“An array of structured activities that represent actual or potential situations in education and practice. These activities allow participants to develop or enhance their knowledge, skills, and attitudes, or to analyse and respond to realistic situations in a simulated environment” (Pilcher et al., 2012, Lioce et al., 2020).

Although healthcare simulations have existed since the 18th century (Carty, 2010), Bradley explains (Bradley, 2006) how the three movements of resuscitation standardisation, anaesthetic simulation and medical education reform drove adoption in the latter 20th century. Other factors

promoting simulation include: the recognition of the frequency and consequences of medical error encapsulated in the seminal report *“To Err is Human: Building a Safer Health System”* (Donaldson et al., 2000), the potential of simulation to prepare for crisis situations more widely than resuscitation (Gaba et al., 2016), a reduction in junior doctors’ hours of training (Romanchuk, 2004), and a changing medical culture that began to view novices practicing on real patients without having first undergone simulator training as unethical (Ziv et al., 2003).

Simulation allows learners to gain proficiency before practicing on patients and increase their competence and self-efficacy with guided repetition. It can help learners to create clear task-related goals and create a ‘safe’ space to make mistakes, as well as facilitating real-time augmented feedback (Norman et al., 2018).

3.4.2 Validation in health professions education

Medicine is at the forefront of simulation research, dissemination and best practice - a large contribution to the validation of simulation in education has been made by data from medical education. In 2011 Cook et al. conducted a systematic review and quantitative meta-analysis (Cook et al., 2011) of 609 studies comparing technology-enhanced simulation with no instruction, of which 137 were randomised. The authors used a broad definition of technological enhancement, including *“computer-based virtual reality simulators, high-fidelity and static mannequins, plastic models, live animals, inert animal products and human cadavers”*. Heterogeneity between the studies was large, but overall large effect sizes^{viii} were seen for the impact of simulation on skills and behaviour outcomes ($d = 1.09$ for process-related - as opposed to time-related - skills; $d = 0.81$ for non-time behaviours); effect sizes that are consistent with the strongest educational interventions (Hattie, 2009). A moderate impact on patient outcomes was seen ($d=0.50$, 95% confidence interval (CI) 0.34 - 0.66) for the small proportion of studies ($n=32$; 5%) that collected patient outcome data. The validity of calculating effect size over such a broad range of studies is open to question but can also be seen as evidence of generalisability of simulation effectiveness over multiple learner groups, instructional designs, clinical topics and contexts, as the authors themselves argue.

The same group compared technology-enhanced simulation with other instructional methods - a more realistic comparator - in a 2012 paper (Cook et al., 2012): in the 92 studies simulation

^{viii} For a helpful and frequently cited primer on effect sizes in education see:
<http://www.leeds.ac.uk/educol/documents/00002182.htm>

improved skills and behaviour (and patient effects for those studies examining patient outcomes) with small to moderate effect sizes when compared with control non-simulated interventions.

Also in 2011, McGaghie et al. published a systematic review of studies specifically comparing deliberate practice simulation training (see Section 3.6 below) versus traditional education or pre-intervention baseline measure (McGaghie et al., 2011). This analysed 14 randomised-controlled or comparative effectiveness trials with a total of 633 learners (a mixture of post-graduate trainees (n=389) and medical students), learning medical or surgical procedural skills. Deliberate practice had a consistent, large beneficial effect ($d=0.71$) for simulated or real-life skill acquisition.

In one example Barsuk et al. trained a cohort of residents (specialist trainees) working in an intensive care unit (ICU) in a single institution using a simulation-based learning programme (Barsuk et al., 2009) which involved a lecture and demonstration followed by at least 3 hours of mannequin-based simulator training and feedback. Residents were required to pass a post-session skills test. Catheter related bloodstream infections were compared with a control ICU in a separate institution with a slightly higher baseline rate (pre-intervention unit 3.2 infections / 1000 catheter-days vs 4.9 in control). Catheter-related bloodstream infections dropped sharply in the intervention ICU whereas they stayed elevated in the control ICU (0.5 vs 5.26, $p = 0.001$) over the 12-month study period.

In a further meta-analysis in 2014 titled *“How much evidence does it take? ...”* (Cook, 2014), Cook argued that “standards of evidence have been met” for the effectiveness of simulation-based education, and that researchers needed to stop asking “does simulation work” to “what works, in what circumstances, and for whom?” (Wong et al., 2010).

It is important to note that despite several simulation studies showing benefits to clinical outcomes, these studies still generally fail to reach the standards of evidence expected for other medical interventions i.e. patient survival and quality of life. Potential reasons for this have been discussed in detail elsewhere (Brydges et al., 2015).

3.4.3 Simulation - what works?

Principles of ‘what works’ were first outlined in a best evidence in medical education (“BEME”) systematic review by Issenberg et al. in 2005 (Issenberg et al., 2005) and have generally been borne out by subsequent reviews (Cook et al., 2013, McGaghie et al., 2010). Simulation design

features whose effectiveness generally holds when implemented across medical specialities and contexts are collated below:

- Curriculum integration, with a range of learning activities / strategies
- Cognitive interactivity
- Range of difficulty
- Representative of clinical variation
- Feedback
- Distributed practice
- Repetitive practice
- Individualised practice

Figure 3-3 - Effective instructional design features in medical simulation. Collated from (Issenberg et al., 2005, Cook et al., 2013, McGaghie et al., 2010).

One aspect of simulation that was initially taken for granted was “fidelity”. The concept is difficult to define operationally (Hamstra et al., 2014) but can be thought of as *“the degree to which the simulation replicates the real event and/or workplace; this includes physical, psychological, and environmental elements”* (Lioce et al., 2020).

Educators initially assumed intuitively that high-fidelity simulations would result in greater learning and transfer (Issenberg et al., 2005). A number of studies show lower-fidelity simulation produces similar skill improvement, often at substantially lower cost. For example, Anastakis et al. randomised 23 surgical residents to cadaveric (high-fidelity), bench model (low-fidelity - reusable [non-human] materials) and textbook-based training for six surgical skills (Anastakis et al., 1999). Adjusted for skill difficulty and trainee experience, the low-fidelity group scored similarly to the high-fidelity group on cadaveric post-training examination (both scored significantly higher than the textbook group). The study was underpowered to detect small differences. Chandra et al. (Chandra et al., 2008) reported similar findings for anaesthetic assistants, with a virtual-reality intubation simulator training adding nothing to skill improvement compared with training using a simple physical model when intubation was subsequently assessed on real patients.

These findings have been echoed across a variety of practical skills as well as other domains such as clinical reasoning and crisis management (Norman et al., 2012, Brydges et al., 2010), especially in novice learners. They are often explained with reference to cognitive load theory (see Section

3.5 below). Hamstra et al. (Hamstra et al., 2014) state that physical aspects of fidelity (“realism” which in and of itself is not necessary or sufficient for learning) should be differentiated from psychological fidelity - “functional task alignment” - which does relate to learning gains.

Questions to advance our current understanding of medical simulation have recently been reviewed by Henriksen et al. (Henriksen et al., 2018) and include: “what makes effective simulation feedback?”, “how should we best configure the type and timing of practice?”, and “how can we match simulation fidelity to the level of the learner?”.

Even using low-fidelity simulators, simulation programmes involve increased resources when compared to non-simulation programmes - therefore effectiveness must increase accordingly. However less than 10% of studies report any cost information and less than 2% report cost comparisons with other approaches (Zendejas et al., 2013), making cost-effectiveness analyses for simulation challenging.

3.4.4 Application to radiotherapy contouring

The barriers to uptake of technology-enhanced simulation in radiotherapy treatment planning are relatively low - these skills were already conducted by computer in the 1990s and early 2000s. Therefore uptake of simulation in radiotherapy contouring (Tai et al., 2002) and dose planning (de Almeida et al., 2002) happened relatively early: simulation involved using existing radiotherapy software and standardised cases, and structuring an educational intervention around them.

A recent systematic review of simulation in radiotherapy education (Rooney et al., 2018) notes that most studies (45/54 = 83%) did not explicitly identify their approach as “simulation”. The majority of included studies related to radiotherapy contouring (54%) or treatment planning (20%), and objective outcome measures were only published for contouring studies - but these were nearly all limited to conformity indices. The authors did not comment on the limitations of contouring simulation discussed in the previous chapter and by Cacicedo et al. despite significant overlap of the studies cited - their focus was on highlighting the “vast potential space” for simulation in radiotherapy education outside of the contouring domain.

The research from simulation in medical education presented above sheds further light on the limitations of contouring simulation so far in radiotherapy. I have already highlighted (see Section 2.4.4) that radiotherapy studies often fail to demonstrate either skill retention (only immediate post-intervention testing is conducted) or transfer (to a different case), and that

distributed or ‘spaced’ practice, which is well-known to facilitate long-term retention (Cecilio-Fernandes et al., 2018), is rarely reported. Lower fidelity simulation may work as well, or better, for novices and the resources and software complexity required can be significantly lower; again this approach has not been studied in radiotherapy and merits consideration. Other features of effective simulation design such as understanding case difficulty and learning curves, incorporating clinical variation, and individualised practice are likely to be part of workplace-based training, but have not been reported in contouring educational interventions. Feedback is part of radiotherapy quality assurance, but seems to be unstandardised (for further discussion see Section 3.8).

Measuring the impact of physician contouring education on patient outcomes (as seen in the surgical specialities) has not been attempted. Although admittedly challenging, arguably radiation oncologists can and should aspire to a greater level of educational impact (Kirkpatrick, 1967, McGaghie et al., 2014) than conformity index on the same case. The aim of contouring education should be to avoid marginal or out-of-field tumour recurrences and thereby increase loco-regional control. Because the impact of complex interventions on hard outcomes such as tumour control is likely to be modest (Ivers et al., 2012) large numbers of patients within a prospective registry would be required to test the impact on patient outcomes - the resources required may be prohibitive given the paucity of funding available for educational research (Reed et al., 2007). A less ambitious goal would be to show improvement in expert-assessed protocol deviations in real cases over time - which would reach the third tier of Kirkpatrick’s levels of educational evaluation:

Table 3-1 - Kirkpatrick’s hierarchy of training evaluation applied to medical education. Adapted from Sharma et al. (Sharma et al., 2015) p.116

Level 1	Learner reaction: satisfaction, perceived improvement in skills and/or knowledge
Level 2	Effect on learner knowledge (2a) or skills (2b)
Level 3	Effect on behaviour in the workplace e.g. increased quality or reduced errors
Level 4	Effect on patient outcomes

In addition, radiotherapy has not moved beyond the question of ‘does simulation work?’, to focus on: ‘what type of simulation works and for what type of learner?’, ‘how should it best be structured and delivered?’, and ‘at what stage of training?’.

3.5 Cognitive load theory

3.5.1 Definition & context

Cognitive load theory was formulated by Sweller et al. (Sweller, 1988) who built on well-validated insights into the limits of human working memory (Miller, 1956) and cognitive architecture (Atkinson and Shiffrin, 1968):

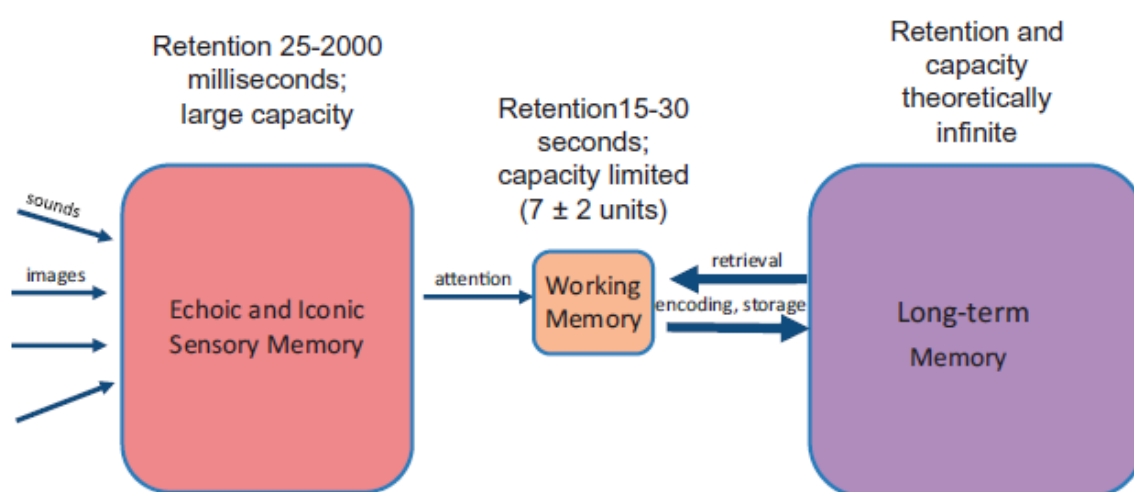


Figure 3-4 - The Atkinson-Shiffrin model of human memory & information processing. Reproduced from Young et al. (Young et al., 2014), p.373

Information enters via the sensory organs and is stored for a very short time (less than 2 seconds) in auditory (echoic) and visual (sensory) memory, whose capacity are somewhat independent of each other. The learner's attention then directs storage of selected information into working memory. This can be processed alongside information retrieved from long-term memory, and/or "encoded" (translated and sent) to long-term memory within knowledge structures known as 'schemata'. The core constructs of cognitive load theory are defined in Table 3-2.

Table 3-2 - Key constructs in cognitive load theory and their definitions

Concept	Definition
Working memory	A limited-capacity system in the human brain that can briefly store (~30 seconds) and process units of information
Schemata (singular = 'schema')	Structures in long-term memory which enable 'chunking' of complex information (such as diagnostic features of a condition, or patterns of tumour spread) into single units. Advanced schema formation allows experts to process higher complexity information concurrently compared to novices.

Intrinsic load	Working memory capacity taken up by features intrinsic to the learning task. Arises from the interaction between a person's expertise and the complexity of a learning task.
Element interactivity	Cognitive load theory construct used to express complexity. Cognitive load increases as information elements interact with each other, and is higher with parallel as opposed to sequential processing.
Extraneous load	Working memory capacity taken up by features related to but not central to the learning task e.g. familiarising self with simulator
Germane load / processing	Working memory capacity or mental effort consumed by processes relating to learning i.e. constructing or 'encoding' schema to long-term memory. Can be considered a component of intrinsic load.

Learning tasks contain information that is essential to complete the task which may include, for example, instructions and case information. Processing these elements puts **intrinsic load** on working memory. In addition, the task may involve processing other visual and auditory information, such as using unfamiliar software - the impact of this on working memory is termed **extraneous load**. Cognitive load theory postulates that the addition of these two burdens on cognitive capacity may cause cognitive overload and reduce learning, whereas reducing extraneous load and optimising intrinsic load leaves capacity for “germane processing” or learning.

Cognitive load theory is increasingly applied to health-professions education (Sewell et al., 2019) with especial relevance for novice learners, whose lack of well-developed memory schemata for information processing make them vulnerable to cognitive overload. The emphasis of instructional design informed by cognitive load theory is on explicit structure and guidance for the learner (Kirschner et al., 2006) as opposed to minimal guidance approaches such as discovery learning (Alfieri et al., 2011).

The expertise reversal effect

The “expertise reversal effect” occurs when the effectiveness of instructional strategies that are known to help novices (such as element isolation and worked examples) decreases with increasing learner experience. This sometimes occurs to the point where they actively hinder learning compared with unassisted practice (Kalyuga et al., 2003, Kalyuga, 2011). A similar phenomenon is seen in the provision of feedback to experienced learners (see Section 3.8.2).

3.5.2 Validation in health professions education

A recent scoping review examined studies of cognitive load in workplace training (including the simulated environment) identified 116 studies relevant to health professions education (Sewell et al., 2019). The review found a profusion of data supporting a positive relationship between cognitive load and simulated task complexity and inverse relationships between cognitive load and prior experience and initial skill performance. These links were most pronounced amongst novices. No studies of cognitive load in radiotherapy have been conducted, but multiple surgical and endoscopic training studies demonstrate this link. For example, in a group of 14 surgical residents and fellows, increasingly complex laparoscopic technique was associated with increased reported cognitive load and worse performance during a simulated task (Montero et al., 2011). Bharathan et al. compared proficiency of gynaecology trainees and experienced clinicians on a virtual reality laparoscopic simulator. The experienced clinicians took less time with fewer movements and reduced reported cognitive load, a finding which is typical in the literature - the relationship between reduced cognitive load and both greater clinician experience and reduced task complexity is “incontrovertible” (Sewell et al., 2019, p.261). These findings fit with those of studies investigating the effects of low-fidelity simulation described above.

The authors of the scoping review are careful to note that most of the 116 studies measured short-term outcomes and generally lacked evidence for transfer to authentic settings - the effect of cognitive-load informed instruction on these outcomes needs to be tested prospectively. In one example of this, Haji et al. randomised 38 medical students to train on simple or complex lumbar puncture tasks (Haji et al., 2016). Self-reported cognitive load decreased during training in the ‘simple task’ group who performed better initially, but not in the ‘complex task’ group. At retention and transfer 10 days later, both groups performed similarly, although there were less breaches of sterility in the ‘simple task’ group.

Measurement of cognitive load is challenging - it cannot be measured directly. Its measurement in medical education has been reviewed by Naismith et al. (Naismith and Cavalcanti, 2015). The most common method of measurement is self-reported cognitive load, whose validity is supported by its relationship with other outcomes such as skill performance. However, it has limitations as it is subject to recall bias and is also difficult to evaluate in sub-tasks without disrupting the learning activity. Another common measurement method is secondary tasks (such as rhythmic foot-tapping (Park and Brünken, 2015)), but this is problematic in evaluating learning as it introduces further load, reducing attention to the task. Physiological measures, for example pupil dilation, learner gaze analysis, and electro-encephalogram (EEG) readings are

promising but are still being validated (Sweller et al., 2019) and at present require expensive technology.

3.5.3 Critique

Some core tenets of cognitive load theory are untestable, or at least very difficult to validate, as cognitive load or overload and schema formation cannot be measured directly (Murphy et al., 2016). Until recently, few attempts have been made to measure sub-types of cognitive load, although self-reported sub-types have been validated in at least one study in health professions skills training (Sewell et al., 2016). This fundamental inaccessibility of the core constructs threatens the validity of this scientific theory and means that explanation of unexpected or contradictory findings can become tautological, making it difficult to advance the theory in light of new evidence (Moreno, 2009).

The construct of germane load has been questioned (de Jong, 2009) - it may be re-framed as a facet of intrinsic load (Sweller et al., 2019), or even extrinsic load if the mental effort required for learning exceeds working memory limitations. The effectiveness of an activity in promoting learning goes far beyond the narrow focus of cognitive load theory and so must be approached with multiple theoretical perspectives - one cannot neglect the importance of learners' affect and motivation, for example (Moreno, 2009).

Categorisation of cognitive load (i.e. extrinsic versus intrinsic or germane) also depends on the learner. Managing distractions whilst suturing in the operating theatre is a peripheral skill for the novice, but is vitally important for a qualified surgeon who will inevitably have to manage both. The medical workplace environment is inherently complex and physicians need to train with that in mind.

Another criticism levelled at cognitive load theory and its instructional design implications is that it offers little new with reference to other established principles of good pedagogy arising outside of the theory. De Jong (de Jong, 2009) highlights that the core design principles - aligning instructional material with the knowledge of the learner, minimising unnecessary information, and stimulating learning processes - pre-date and are independent of cognitive load theory. For pragmatic purposes however, the alignment of these design principles is *reassuring* in that instruction based on them is likely to be pedagogically sound.

3.5.4 Application to radiotherapy contouring

Cognitive load theory provides an explanation for the positive effects of reduced simulation fidelity on novice learners. As mentioned above, this approach has not been studied in radiotherapy contouring. The established principle that novice and experienced learners process information in different ways is also highly relevant, as generally in radiotherapy contouring instructional design has been similarly conceived for both novices (junior clinical/radiation oncology trainees) and accredited clinicians.

Cognitive load theory provides further insight to the solutions and problems seen in the previous chapter, for example the effectiveness of worked examples (radiotherapy atlases) and the problems stemming from lack of practice variability or sequencing in contouring training.

The design principles developed from cognitive load theory have been summarised by Young et al. (Young et al., 2014) are presented in Table 3-3 below together with my interpretation of their potential relevance to radiotherapy contouring. Whilst some have already been addressed with the introduction of contouring atlases, others including the sequencing of instruction and matching the level to the learner, require innovation.

Table 3-3 - Application of cognitive load theory to instructional design in radiotherapy contouring. Adapted from Young et al. (2014) pp. 380-2

Design principle	Instructional technique	Potential application to radiotherapy contouring
Decrease extraneous load	Worked example - provide learner with a demonstration of the problem solution	Already implemented in radiotherapy atlases, however atlases are <i>product</i> rather than <i>process</i> examples which are more effective for experts than novices (van Gog et al., 2008)
	Problem completion - provide learner with a partially completed problem and ask them to complete the next steps	Part-task training (e.g. providing the gross tumour volume and asking trainees to expand to the clinical target volume). This has not been attempted.
	Avoid split attention - present instructional diagrams and text together	Synthesize contouring atlases to group the instructions with the illustrations
	Use multiple sensory modalities (auditory and visual)	Provide auditory commentary with worked examples
	Allow learners to refer back to transient information	Already implemented with written instructions in contouring guidance
	Avoid redundant information	Break up contouring into steps with learning material provided just in time
Manage intrinsic load	Isolate elements of information	Break up contouring into steps to avoid cognitive overload
	Pre-training	A lecture or educational module is already commonly provided prior to contouring practice in radiotherapy
	Progress from low- to high physical fidelity	Requires the development of a low-fidelity contouring simulation
	Progress from simple to complex	Requires classification of contouring tasks by complexity and then content creation to allow progression. Current radiotherapy contouring training relies on random case mix in learner rotation
<i>Optimise germane processing</i>	Contextual interference (i.e. mixed rather than blocked practice) & variability of practice	Create multiple examples of same principle at each level of difficulty
	Self-explanation - encourage learners to explain a concept or learning task to themselves	This could be incorporated into contouring practice routines once a programme of education is established
<i>Expertise reversal</i>	“Progressive completion” or reduction in scaffolding - reduce learner support as experience increases	Reduce level of instruction and feedback as contouring competency increases

3.6 Deliberate practice theory

3.6.1 Definition & context

The theory of deliberate practice was outlined by Ericsson et al. in their seminal research on elite performers in music, chess & sport (Ericsson et al., 1993, Ericsson and Charness, 1994). Their first study compared the self-reported practice habits of the highest-performing violinists in a Berlin Music Academy with those of their less accomplished colleagues and those of aspiring music teachers from a different institution. These data were then contrasted with similar data provided by older professional violinists. Figure 3-5 below shows the estimated accumulated hours of practice for each cohort:

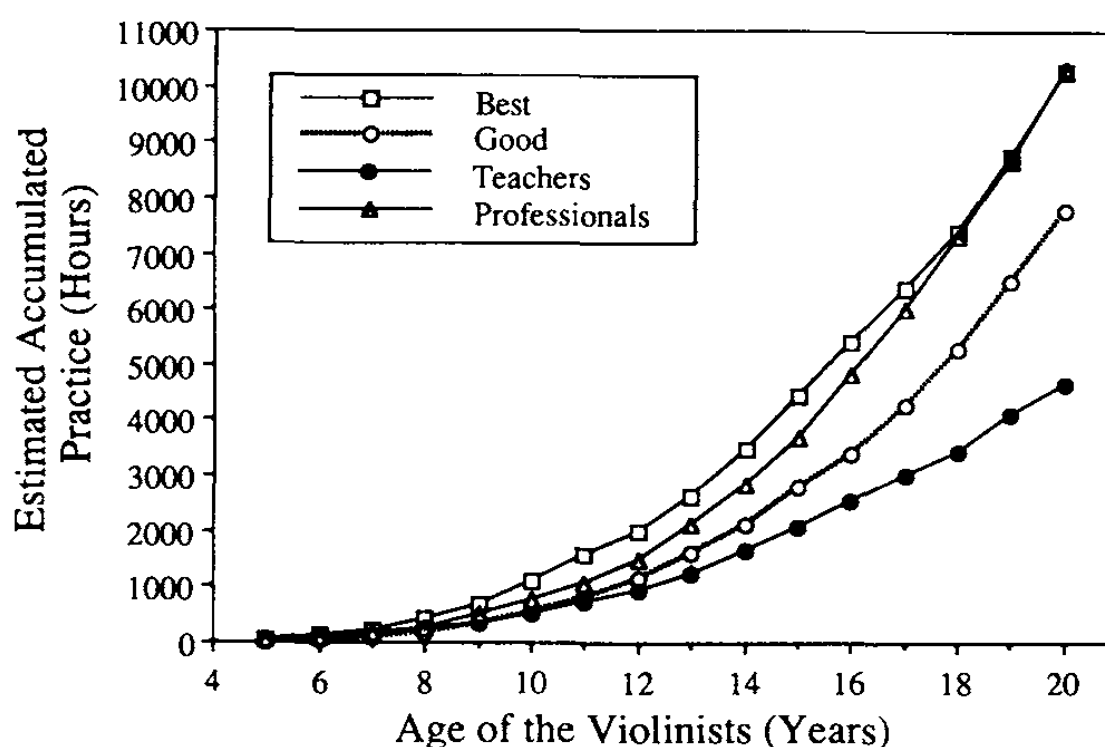


Figure 3-5 - Estimated accumulated practice hours of four cohorts of Berlin violinists in Ericsson et al.'s seminal work on deliberate practice. Reproduced from Ericsson et al., 1993

The authors' conclusion from this study was that the variance in performance between the different cohorts was almost entirely explained by accumulated hours of practice, and did not "depend on scarcity of innate ability (talent)" (Ericsson et al., 1993, p.393). Ericsson subsequently defined deliberate practice as:

"the individualized training activities specifically designed by a coach or teacher to improve specific aspects of an individual's performance through repetition and successive refinement" (Ericsson and

Lehmann, 1996, p.278-9). This results in gradual development of the individual's abilities to plan, execute and self-monitor their performance (Figure 3-6), and is “effortful”, limiting the amount of time that can be spent on deliberate practice per day (Ericsson, 2015).

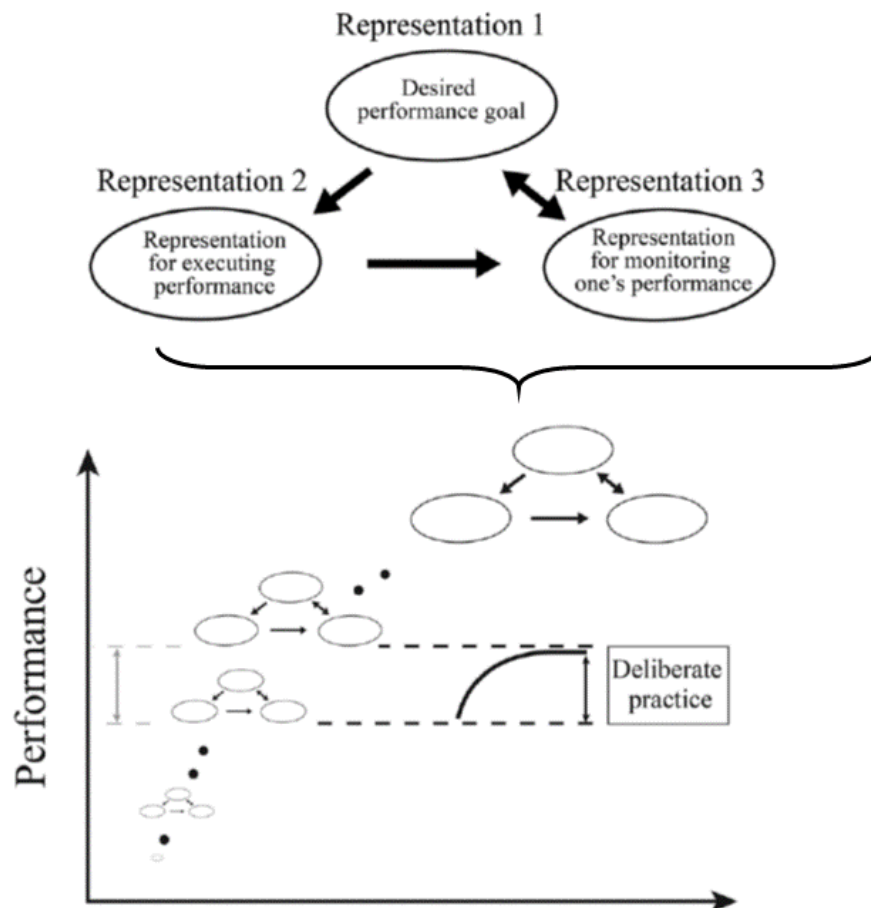


Figure 3-6 - Schematic illustration of deliberate practice as a journey towards expertise with increasingly sophisticated performance representations. Reproduced from Ericsson, 2015 (p.1473 & 4)

This research has had a huge influence in the science of expertise (Macnamara and Maitra, 2019) as well as in the popular imagination; several bestselling books are based on their findings (for example ‘Outliers’ by Malcom Gladwell and ‘Bounce’ by Matthew Syed) which are the foundation for the popular claim that it takes “10,000 hours” of practice to become an expert.

Deliberate practice theory has been enthusiastically adopted by medical education researchers, especially for practical surgical skills after the reduction of training hours in the 2000’s (Higgins et al., 2020). Deliberate practice for clinical skills has often been paired with a ‘mastery’ learning approach although Ericsson himself has highlighted that the two constructs, though overlapping, are different (Ericsson, 2015).

McGaghie (McGaghie et al., 2009, McGaghie, 2015) has defined deliberate practice within a mastery learning programme as possessing at least the following seven features:

Table 3-4 - Features of mastery learning. Reproduced from McGaghie, 2009, p.65S

- | |
|---|
| <ol style="list-style-type: none"> 1. Baseline or diagnostic testing 2. Clear learning objectives, sequenced as units in increasing difficulty 3. Engagement in educational activities (e.g. skills practice) focused on reaching the objectives 4. A set minimum passing standard for each educational unit 5. Formative testing to gauge unit completion at a present minimum standard for mastery 6. Advancement to the next educational unit given measured achievement at or above the mastery standard 7. Continued practice or study on an educational unit until the mastery standard is reached |
|---|

3.6.2 Validation in health professions education

Multiple studies validate the role of simulated deliberate practice as superior to traditional teaching when examining skill improvement - McGaghie's meta-analysis of 13 studies was mentioned above (McGaghie et al., 2011). There is also evidence of improved patient outcomes resulting from deliberate practice simulation training (Higgins et al., 2020). In one example Zendejas et al. randomised 50 surgical trainees to standard simulated practice of inguinal hernia repair versus a mastery learning curriculum with deliberate practice (Zendejas et al., 2011). The mastery programme consisted of 9 online multimedia modules assessed with a multiple-choice knowledge test, followed by supervised practice with a laparoscopic task trainer with a requirement to achieve simulator performance comparable to experienced surgeons. Trainees in the control arm took part in self-directed learning and intraoperative learning i.e. they did not participate in simulation training. The deliberate practice with mastery learning cohort demonstrated faster operative times and were scored higher in blinded assessments, and in addition their patients less frequently stayed overnight (0/48 patients versus 9/38, $p < 0.001$) and suffered fewer intraoperative complications (7% of patients versus 29%, $p = 0.03$).

The effects of domain-specific practice are corroborated by data correlating surgical experience with cancer outcomes (Vickers et al., 2007) although the distinction between deliberate practice and simple 'practice' is not clear in this setting. Clinicians' number of years of medical experience beyond the completion of training do not correlate with performance (Choudhry et al., 2005), which implies that there is a 'quality' of practice which is required for performance improvement.

What is clear is that expertise in one specific procedure does not necessarily translate into another type of procedure in a related domain. This was demonstrated in the transition from open to laparoscopic surgery in the 1990s-2000s, where experience with the laparoscopic procedure was shown to be significantly more important in determining outcome than surgical experience with open procedures (Moore and Bennett, 1995). Clinicians who are changing their practice may need to be trained in a similar way to novices (Norman et al., 2018).

Both deliberate practice theory and cognitive load theory are supported by findings from studies examining the effect of progressively increasing simulation exercise difficulty: in one example Grover et al. randomised 37 novice endoscopists on a simulation programme to practicing progressively more difficult exercises or practicing exercises in a random order (Grover et al., 2017). Observed performance on their first two clinical endoscopies was higher for the 'progressive learning' group. Progressively increasing simulation fidelity has shown similar effects (Brydges et al., 2010).

High performance *during* practice does not necessarily correlate with performance when testing for retention and transfer (Schmidt and Bjork (1992)); in fact difficult learning exercises can suppress immediate performance during practice but result in greater long-term learning gains - the optimal practice difficulty depends on the ability of the learner. This well-validated phenomenon is modelled by the challenge point framework (Guadagnoli et al., 2012). The necessity of challenge is also emphasised in Vygotsky's 'zone of proximal development' - skills that are at the limit of a student's competence, achievable with instruction (Vygotsky, 1962, Konopasek et al., 2016).

3.6.3 Critique

Initial claims by the researchers who formulated deliberate practice theory, for example that *"it is possible to account for the development of elite performance among healthy children without recourse to unique talent (genetic endowment) excepting the innate determinants of body size"* (Ericsson, 2007a, p.199) are clearly overstated (Ackerman, 2014). Gardner argues this statement "requires a blindness to ordinary experience" (Gardner, 1995, p.802) where individuals clearly differ in their ability to perform and progress in skills such as mathematics, writing essays or learning a sport or musical instrument. A meta-analysis (Macnamara et al., 2014) of the relationship between performance and accumulated practice showed that time spent practicing explained the **minority** of variance in performance in the domains of computer gaming (26%), music (21%), sports (18%) and education (5%), although in the former three domains it was clearly an important factor. Only 1 study in medical education was included in the latter domain.

Macnamara et al. recently attempted to replicate Ericsson's original research study (Macnamara and Maitra, 2019) in a US cohort. They found a smaller effect size of practice and that intermediate-performing group reported having completed *more* practice than the most accomplished performers:

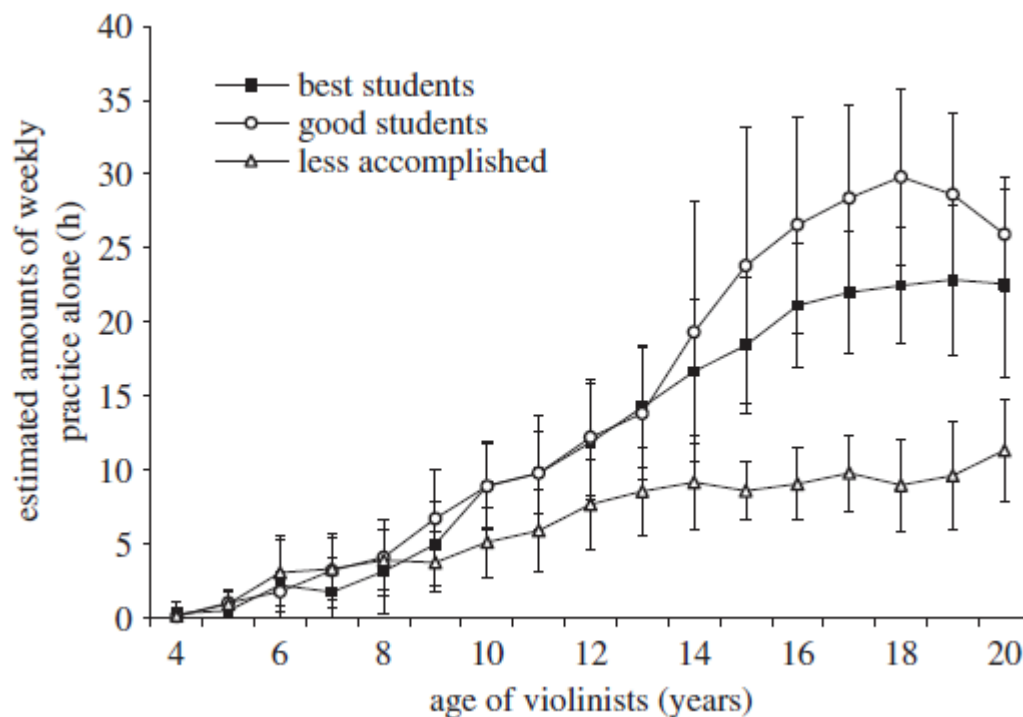


Figure 3-7 - Practice hours as a function of age in Mcnamara et al.'s study attempting to replicate Ericsson et al.'s methods

Replication of seminal psychology studies is known to be problematic - Macnamara et al.'s findings are not surprising given the results of efforts by the Open Science Collaboration to replicate 100 high-impact psychological studies (Open Science Collaboration, 2015).

'Deliberate practice' can also be difficult to distinguish from simple 'practice' - various definitions have been used; sometimes only including only teacher-designed activities (Ericsson, 2015) but on other occasions also including learner-designed practice (Ericsson, 2007b).

Ericsson himself has noted that reliably superior performance is difficult to identify in medicine when compared with the domains of sports, music and chess. However, it is undeniable that practice has a significant role to play in skill development (Macnamara et al., 2014) - look again at Figure 3-7: the difference in accumulated practice between the 'less accomplished' and both the 'good' and 'best' performers is clear. 'Good' may be good enough for the purposes of practical skills training in medicine.

3.6.4 Application to radiotherapy contouring

There is strong evidence of the effect of the amount and quality of deliberate practice on performance. Acknowledging that other factors influence performance does not nullify the large effect of practice & potential for learning gains. Arguably we can aim to help learners journey from novice or 'less accomplished' to 'good' (the transition where the effects of practice are more clearly demonstrated) rather than 'best' or 'elite' performance and still expect to reap benefits to radiotherapy quality.

Standards for mastery (or competency) need to be established for radiotherapy contouring to judge progression along a learning curve. This is a major challenge given the variation between experts. It may be easier for straightforward or common rather than complex or unusual cases (Weiss et al., 2003). There is work to be done to identify and verify what tasks are easy and difficult for each tumour site. Contouring organs at risk is generally more straightforward than target volumes and this is often where trainees begin contouring, however there are no published programmes of contouring mastery with progressive difficulty such as those developed in surgery and endoscopy. Progressive simulation fidelity adapted to the learner is possible but would require the development of a low-fidelity simulation with associated content. Shared content could facilitate accessibility as current contouring courses are relatively limited in content (to one or two cases) and can be expensive.

In contrast to surgical training and practice (The Royal College of Surgeons of Edinburgh, 2021), there are no requirements to keep logbooks of radiotherapy procedures in the UK, or minimum number of cases required for competency. In a different vein, surgeons in UK are ahead of clinical oncologists in publicly available outcome data (Radford et al., 2015, Royal College of Surgeons of England, 2014). Catching up with our surgical colleagues - at least detailing the aims, hours and achievements of simulated practice and/or clinical experience - is an achievable aim that the radiotherapy community should work towards. Simulated deliberate practice is an important part of this journey towards mastery, but new training standards, structures and ideally shared curricula and technologies will have to be developed (Evans et al., 2019b).

The data from surgical changes in practice and expertise have implications for clinicians changing their radiotherapy contouring approach (especially, one suspects, if this involves a change in target concepts). Instructional techniques that work for novices such as worked examples and repetitive practice (potentially using low-fidelity simulation) should be considered - the former can be seen in studies of radiotherapy atlases (Eminowicz et al., 2016a) but the latter has mainly been employed in interventions limited in the duration and scope of practice (see Section 2.4.4).

3.7 Assessment

3.7.1 Definition & context

Assessment is defined in *Standards for Educational and Psychological Testing* as “a systematic process to measure or evaluate the characteristics or performance of individuals, programs, or other entities, for purposes of drawing inferences” (American Educational Research Association, 2014, p.216).

The requirement to certify health professionals as capable of delivering patient care means that the medical education literature on assessment is well-developed and extensive. This includes simulation-based assessments (Barrows, 1968, Hatala et al., 2005, Brydges et al., 2015).

‘Miller’s pyramid’ (Miller, 1990) is a foundational lens through which to view assessment of doctors’ competency and illustrates different levels of learning and assessment:

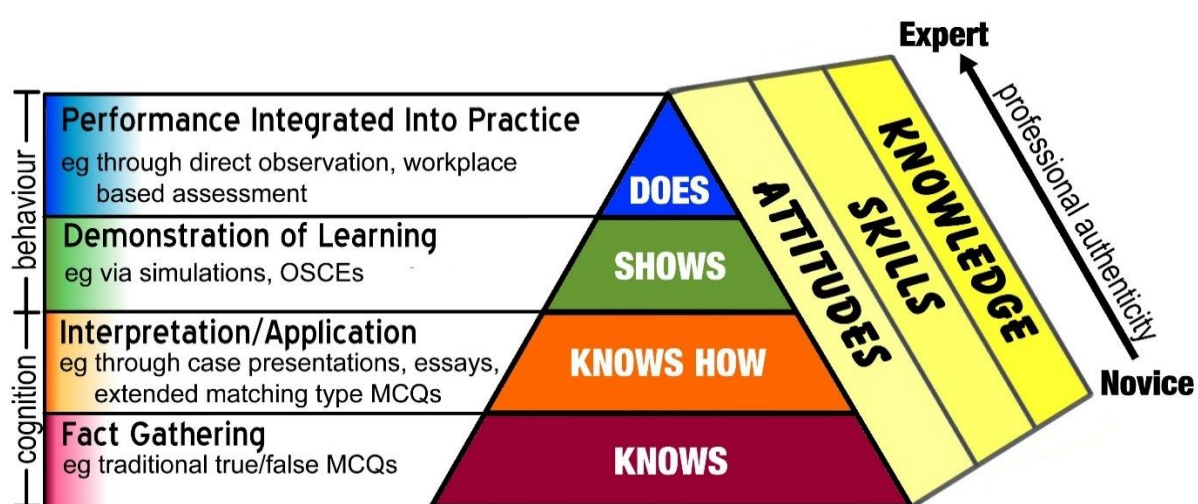


Figure 3-8 - Miller's pyramid. Adapted from original work by Miller (1990) by Mehay & Burns (Mehay and Burns, 2009)

Assessments can also be categorised by their format (Yudkowsky et al., 2019): written tests, oral examinations, performance tests and workplace-based assessment. This section will focus on assessment of performance as this is most relevant to skills simulation.

‘Competency’ is a key concept in assessment and is defined as “the degree to which the individual can use the knowledge, skills and judgement associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice” (Kane, 1992p. 166). Competency-based education (including the concept of ‘entrustable professional activities’ (ten

Cate, 2005)) is the foundation of recent post-graduate curricular reform (Nasca et al., 2012), including in Clinical Oncology (Royal College of Radiologists, 2020a). Valid assessments are a key part of any competency-based programme.

Whereas assessments of competency with traditional standard setting processes aim to identify the ‘just-competent’ or ‘just-safe’ candidate, mastery assessments seek to ascertain that the learner has achieved a certain *completeness* of knowledge or skill (Lineberry et al., 2015) and that “*all learners are well prepared to succeed in subsequent stages of training*” (Yudkowsky et al., 2015, p.1495).

Other key assessment terms for this thesis are defined in Table 3-5:

Table 3-5 - Key terms in assessment relevant to this thesis

Term	Definition
Summative	Attempts to measure final ‘achievement’ in a topic, domain, or skill. Often for purposes of certification. Can be “high-stakes” or “low stakes” depending on consequences. Situated towards the end of a course of study.
Formative	Designed to provide feedback to the learner and/or teacher. Not generally used for certification. Situated during the course of study.
Norm-referenced	Assessments whose standards are relative to other candidates or cohorts.
Criterion-referenced	Assessments that have a standard relating to knowledge or skill level; pass rate could be between 0 - 100%. E.g. Objective Structured Clinical Examinations in medical school finals.
Compensatory	Candidates can make up for poor performance in one case, station or domain (e.g. a particular tumour site) in another.
Conjunctive	Candidates must pass all aspects or domains to pass the assessment.
Global rating scale	Allocation of a rating based on overall impression rather than specific checklist or sub-task items (e.g. in radiotherapy contouring - ‘good’, ‘minor deviation’, ‘major deviation’).

3.7.2 Assessment utility & best practice

Van der Vleuten listed the important variables determining assessment utility in the assessment of competency utility as: validity, reliability, educational impact, acceptability and cost/feasibility (Van Der Vleuten, 1996). Norcini et al. re-affirmed these in a consensus statement (Norcini et al., 2011) which expanded on validity and educational aspects.

The validity of an assessment is *“the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests”* (American Educational Research Association, 2014, p.11). If an assessment lacks validity for a specific interpretation, its results may be meaningless for that purpose. Sources of validity evidence are detailed in *Standards for Educational and Psychological Testing* and have been applied to medical education (Downing, 2003, Downing and Haladyna, 2004) - they are: content, response process, internal structure, relationship to other variables, and consequences and are outlined in Table 3-6. Within this framework assessment reliability is incorporated as a facet of validity, although in the literature it is sometimes treated separately (Downing, 2004). Their potential application to radiotherapy is also presented in Table 3-6 below. The higher the stakes of the assessment, the greater the amount and monitoring of validity evidence is required (Downing, 2003).

Clearly an assessment that is prohibitively costly in monetary terms or in terms of faculty and/or learner time will not be practical, however valid. As Van der Vleuten states: *“Perfect utility is utopia. In practice we will always be required to compromise ...”*. assessment validity therefore may be limited by resources. The educational impact of assessment, including formative assessment, is discussed in the following section (3.8) on feedback.

Table 3-6 - Sources of assessment validity with potential application to radiotherapy contouring assessments. Adapted from Downing (2003), p.832 and American Educational Research Association (2014), p.11-22

Source of validity	Facet & explanation	Application to radiotherapy contouring assessments
Content	Representativeness of assessment content for intended purpose	Case mix should be representative of clinical practice and/or trial eligibility criteria
Response process	Candidate familiarity with response format	Practice with contouring software and adjuncts (e.g. clinical information) before assessment
	Measurement instrument and associated sub-scales	Scoring checklists - rare in radiotherapy quality assurance assessment. Clinical implications of conformity indices uncertain
Internal structure	Item difficulty / discrimination	No published data for radiotherapy contouring
	Relationship of scores for different responses	No data for radiotherapy contouring - again limited by case mix. Could explore different regions of interest and targets vs organs at risk
<i>reliability</i>	Accuracy and consistency of assessment	Inter-rater reliability hardly explored in radiotherapy contouring expert review. Conformity indices highly reproducible but may lack validity.
<i>generalisability</i>	Evidence that responses generalise to items not tested	Some trials have published results of post-RTQA case review, but generally limited evidence
Relationship to other variables	Internal variables	Evidence of association between anatomical knowledge and contouring performance
	External variables	Strong association between prospective clinical protocol deviations ('global' ratings) and patient outcomes, although direct causation not established (see 2.3.3).
Consequences	Impact of results on candidates & system	Radiotherapy quality assurance exercises often repeated without the need for remediation; trial quality may not be fully assured
	Consequences for future learning	Not explored in radiotherapy

Best practice

Devine et al. outline a 7-stage process for simulated assessment design (Devine et al., 2019, p.210):

- Determine learning outcome(s) (intersection of curriculum and assessment content)
- Choose assessment method (appropriateness of simulated assessment and its validity)
- Choose simulation modality (including fidelity)
- Develop assessment scenario (environment, case(s), flow)
- Score assessment (data collection, rating scale)
- Set standards (see below)
- Standardise test conditions (standardise conditions and examiner rating; piloting)

3.7.3 Application to radiotherapy contouring

The scope of current practice in radiotherapy contouring assessment is outlined in Chapter 2 (see Section 2.4.6) - radiotherapy quality assurance is by far the most common formal summative assessment, but could be considered 'low stakes' as it can often be repeated (on the same case) multiple times to obtain a pass. As mentioned in Chapter 2, manual assessment by single or multiple representatives of the trial management group is the most common method of contour evaluation. This is commonly performed on a global rating scale, for example: 'acceptable', 'minor deviation', and major deviation (Weber et al., 2011, Weber et al., 2012). A single study has reported interobserver reliability for contour assessment (McCarroll et al., 2018): the authors compared the 3-point global ratings of 5 experienced radiation oncologists for 8 auto-segmented organ at risk contours in 10 patients. Clinicians disagreed about the clinical acceptability of the auto-contour in 7% of the assessments (2-20% depending on the structure), and disagreements about the need for minor edits (as opposed to no edits) were common - they were seen in 48% of the assessments. Given the increase in contour variability for target volume contours, one can speculate that inter-observer disagreement would be higher.

Details of contouring 'errors' with or without their clinical consequences have been published for some tumour sites (McLaughlin et al., 2010, Joo et al., 2017). These could form the basis for checklists for more reproducible assessments (Daniels et al., 2014) as part of a formal standard setting process, which has not yet been reported in radiotherapy contouring. They may also be helpful for peer review of contours, which again largely proceeds on the basis of global, unstructured assessments.

Conformity indices are very commonly reported as a method of contouring assessment, however although they can provide some guidance for areas on which to focus manual assessment (Conibear, 2018) they are not yet used as the benchmark for clinical acceptability. They are 'objective' with high (approaching complete) 'reliability', but not necessarily high validity - they have limited evidence for a strong relationship to other variables such as expert assessment or clinical outcomes; they do not take into account the location of the variation and cannot differentiate between systematic and random errors (Valentini et al., 2014). Gautam et al. examined the clinical acceptance rates of manually-generated versus auto-segmented contours: they found that conformity index analysis indicated similar acceptance but that manual assessment rated the manually-generated contours higher (Gautam et al., 2013).

As discussed in Chapter 2, content validity in radiotherapy contouring assessment is sub-optimal- as often only one case is assessed, presumably because of time and/or resource constraints. This lack of comprehensiveness is called "construct underrepresentation" (Devine et al., 2019, p.223-4).

The 'response process' in radiotherapy contouring assessment often has high physical fidelity. Often radiotherapy quality assurance is conducted with clinicians' own software which negates this problem but carries additional burden in terms of accessibility and image transfer logistics which could impair engagement in an educational context. Current educational simulations of radiotherapy contouring have much of the functionality of clinical systems. However if clinicians are unfamiliar with what is complex software this may increase their cognitive load and consequently impair their performance.

The 'contrasting groups' standard setting method examines the difference between the performance of candidates/learners compared to certified practitioners or experts in the field and uses these data to define the 'minimally competent' candidate (McKinley and Norcini, 2013). For radiotherapy contouring standard setting is challenging due to the variation amongst clinicians, including expert groups. Defining areas where there is good agreement between expert clinicians is a possible first step for assessment. Examples could be including the 'core' tumour or lymph node basin(s) (by consensus minimum volume), or excluding an organ at risk not in danger of tumour spread, and by contrast areas of acceptable variation (for example areas of uncertain image interpretation or potential spread) could be left unassessed or simply commented on.

3.8 Feedback

3.8.1 Definition & context

Feedback has been described as *“the cornerstone of effective clinical teaching”* (Cantillon and Sargeant, 2008), and the importance of effective feedback for simulation based medical education is well known (Issenberg et al., 2005, Issenberg et al., 1999). Feedback is a key component of formative assessment (Black and Wiliam, 2009) and is amongst the top 10 highest influences on learning (Hattie, 1999, Ko and Sammons, 2013). Feedback has been defined within medical education as: *“specific information about the comparison between a trainee’s observed performance and a standard, given the intent to improve the trainee’s performance”* (van de Ridder et al., 2008). Feedback is a core feature of effective medical simulation with a moderate to large effect size on skill outcomes (Cook et al., 2013, McGaghie et al., 2011).

Characteristics of feedback relevant to practical skills training, referring to its content and timing, are listed and defined in Table 3-7 below:

Table 3-7 - Characteristics of feedback relevant to practical skills training

Term	Description	Reference
Verification	Denotes whether the learner’s answer is correct or not. May or may not provide correct response.	(Shute, 2008)
Elaboration	Addresses the task or topic “providing relevant clues to guide the learner towards a correct answer”. Can involve explanation, hints, and/or worked examples	(Shute, 2008)
Concurrent	Feedback given during a task or sub-task	(Hatala et al., 2014)
Terminal	Feedback given at the end of a task or sub-task	(Hatala et al., 2014)
Immediate	Feedback that is given immediately after a task or assessment	(Van der Kleij et al., 2015)
Delayed	Feedback that is not delivered immediately after completing each item of a task or assessment	(Van der Kleij et al., 2015)

‘Debriefing’ is a specific type of feedback which usually refers to feedback after in-person group simulations (see McGaghie et al., 2010) - the debriefing literature will not be explored in-depth here.

Multiple models of feedback have been developed within medical education (Sargeant et al., 2015, ten Cate, 2013, van de Ridder et al., 2015). Over time there has been a progression from the behaviourist models (feedback travels linearly from the instructor to the recipient) to more complex representations acknowledging the importance of both the learner and the context in which the feedback is provided.

3.8.2 Insights from general education

Many articles have reviewed components and constructed models of effective feedback, perhaps most notably that of Hattie and Timperley (Hattie and Timperley, 2007). Principles of effective feedback were also outlined by Shute in a narrative meta-review (Shute, 2008), and Lefroy et al. in guidelines for medical education (Lefroy et al., 2015). All of these reviews advise that feedback should be specific to the task, adapted to the learner, and practically actionable. Feedback is not always helpful though. In fact many meta-analyses contain studies where feedback was shown to be harmful - the archetypal meta-analysis is that of Kluger and DeNisi (Kluger and DeNisi, 1996) where over a *third* of studies demonstrated a negative effect of feedback. Detrimental or lessened feedback effects were associated with a lack of elaboration and a focus on the person rather than the task, which is in keeping with attribution theory (Weiner, 1972). Simple praise has attenuating or negative effects, and unsurprisingly feedback which is designed to discourage learners negatively impacts performance.

Feedback does not have the same effect on all types of learners. Experiments with teaching motor tasks have shown that frequent feedback (every task) improved short-term performance during practice, whereas infrequent or delayed feedback (after a batch of tasks) leads to lower short-term performance but higher performance on delayed testing (Schmidt and Bjork, 1992, Guadagnoli et al., 1996) - this effect may be heightened in learners with more experience (Guadagnoli et al., 2012). Potential explanations include that feedback could block information processing activities or that variability in feedback could prevent the development of a stable mental representation of the underlying skill (Schmidt and Bjork, 1992).

A meta-analysis of feedback in computer-based education examined qualities of effective feedback (Van der Kleij et al., 2015). Imparting knowledge of the correct response was more effective than simply verifying whether the learner response was correct or incorrect, (which was ineffective, in keeping with previous literature) and there was a moderately increased effect of elaborative feedback over verification alone. There was no statistically significant interaction of feedback timing (immediate versus delayed), although there was a suggestion of potential benefit

of delayed feedback for 'higher-order' skills where students had to apply their knowledge to a problem.

3.8.3 Validation in health professions education

A meta-review of studies reporting variables affecting the process and outcome of feedback in medical education was conducted by van der Ridder et al. (van de Ridder et al., 2015). They examined 46 studies and divided the variables based on their effect on four phases: observation, task performance, feedback provision and feedback reception.

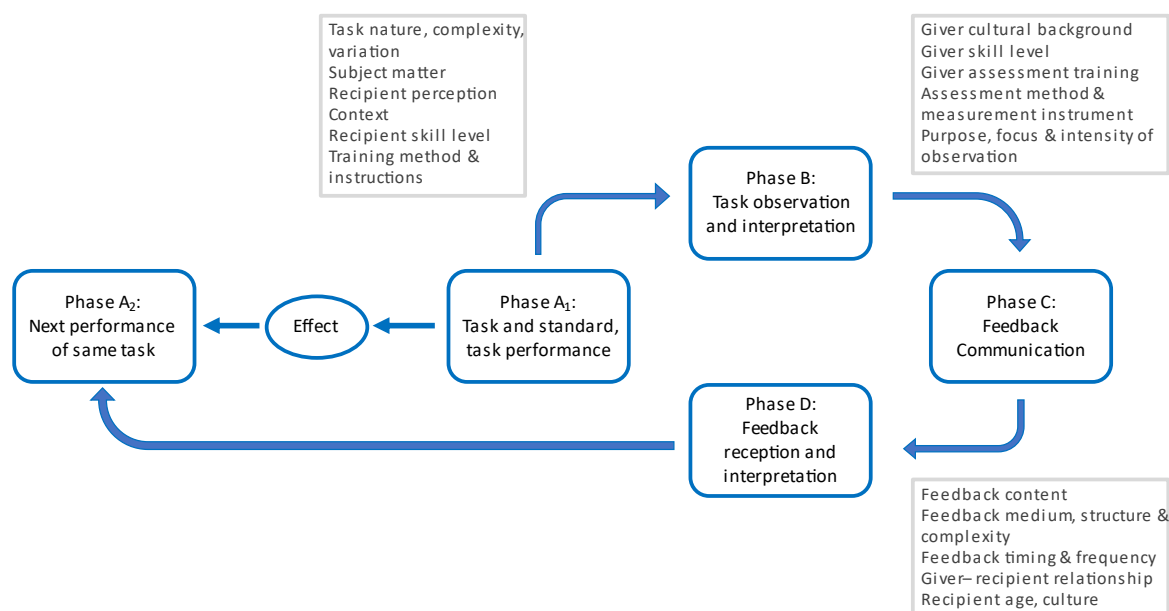


Figure 3-9 - Model of feedback and associated variables. Adapted from van der Ridder et al, 2015

They found that although variables from all phases of feedback were relevant, most studies focussed on the quality of observation and rating of task performance. Variables that corresponded to an unequivocally *increased* effect of feedback were (van de Ridder et al., 2015, p.666):

- Low initial task performance of the feedback recipient
- Message that did not threaten the feedback recipient's self-esteem
- Goal-setting behaviour by the feedback recipient
- Feedback as part of a multi-faceted intervention
- Feedback content that was encouraging, specific and elaborative
- Feedback that was given frequently

Hatala et al. conducted a meta-analysis and realist review of feedback interventions in simulation-based procedural skills training (Hatala et al., 2014). In the 31 studies included, feedback exerted a moderate effect on skill outcomes when compared with no feedback ($d = 0.74$, 95% C.I. 0.38 - 1.09; $P, 0.001$). Terminal feedback appeared to be more effective than concurrent feedback, especially for novices. The authors state that this is consistent with the 'guidance hypothesis' in motor learning where constant feedback leads to an over-reliance on instructors' prompts and results in a slump in performance when the feedback is withdrawn, which is relevant to the concept of scaffolding discussed above (see cognitive load theory Section 3.5.4). As with general educational research, when the post-intervention skills test was delayed or involved transfer then delayed or terminal feedback seemed more effective. One limitation in applying these findings to skills training in post-graduates was that only 5/31 studies were conducted in groups of trainee clinicians or accredited practitioners - most commonly groups of medical students were studied. Hatala et al. also point out that few interventions were grounded in an engagement with theory, so that how the effects of feedback are mediated remains unclear.

3.8.4 Application to radiotherapy contouring

Although there is still much to be explored in feedback for postgraduate skills training, well-established principles of effective feedback can still provide a helpful baseline and are not universally applied in radiotherapy contouring training.

Specific, individualised feedback from contouring courses commonly consists of conformity indices (i.e. a 'score' which may have a normative comparison but which may not denote clinical acceptability), and a gold standard contour ("knowledge of correct response" - although this is somewhat complicated by the disagreement between experts). Groups of contours are often discussed in workshops but resources prohibit individual feedback. There is minimal evidence of individualised elaborative feedback being given in contouring simulation training, other than 1:1 guidance between trainees and their clinical supervisors and in trials as part of RTQA. In these settings, there is no standardised approach and the formulation and delivery of this feedback is time-consuming.

Frequent feedback, another important factor in feedback effectiveness, could be structured as part of a learning programme but as discussed above this would involve an archive of cases or exercises, which would ideally increase in difficulty or complexity in keeping with mastery learning principles. Automated elaborative feedback may retain feedback effectiveness and reduce the burden on supervisors, especially for novice learners whose task performance is likely

to be low. However, this would require technical development of currently available simulations and for experts to agree on feedback (i.e. correct and incorrect responses, and explanations).

The other factors that unequivocally enhance feedback effectiveness (goal setting, situation feedback within a multi-faceted intervention) highlight the importance of structuring the overall educational intervention as a programme of activity relevant to the clinician's practice.

3.9 Summary

This initial exploration of just four domains of educational theory shows the relevance of findings from education research to simulated teaching and assessment of radiotherapy contouring. The most direct application comes from studies of procedural skills training and assessment, but more broadly considering the educational literature also yields valuable insights.

It is clear that we in the field of radiotherapy contouring education can learn much from the simulation literature, deliberate practice theory, cognitive load theory, and established principles of assessment and feedback. Few of these theories have been consciously applied to or tested in radiotherapy contouring education - applying them represents an opportunity to improve educational outcomes.

It is also clear that this hermeneutic review of relevant educational theory is far from comprehensive or complete (see Figure 3-2). During this thesis I will need to engage with other domains of educational theory such as theories of skill development and retention, motivation, self-regulation and metacognition, and human-computer interaction. Nevertheless, the fields outlined and discussed above form a core of relevant theory which can be readily applied to simulated assessment and teaching of radiotherapy contouring.

In Chapter 4 I will move on to describe the theoretical and methodological approaches to answering my research questions, and the structure of the empirical studies in this thesis.

4 Methodology

This chapter outlines the overarching research questions for this programme of doctoral research. I then outline pragmatism as my epistemological paradigm, and design-based research as an appropriate mixed methods methodological framework for the empirical studies that follow. I conclude with an overview of the EMBRACE-II trial in locally advanced cervical cancer, and the rationale for exploring my research questions with practicing radiation oncologists within the EMBRACE-II trial and also with UK- and internationally-based trainee oncologists.

4.1 Overarching research questions

Chapters 2 and 3 proposed that the assessment and teaching of radiotherapy contouring has the potential to be improved by technological and pedagogical innovation founded in established educational theory and best practices. Therefore for this thesis, the research questions asked at the start of Chapter 3 remain central:

- How can medical education literature and the wider educational literature regarding simulated practical skills training inform our approach to the teaching and assessment of radiotherapy contouring?
- How can this knowledge be applied to shape the simulated assessment and teaching of radiotherapy contouring?

During this research I am seeking to develop and evaluate innovative technology-enhanced approaches in this field, and this aim drives the third research question:

- What are the impacts of novel approaches using web-based technology on the teaching and assessment of radiotherapy contouring in the ‘real world’?

4.2 Epistemological considerations

The importance of explicit epistemological assumptions, theoretical frameworks and methodological approach(es) are frequently stated in the field of medical education (Varpio et al., 2020, Sullivan et al., 2014, Reed et al., 2007). The lack of theoretical framework and associated methodological approach is a common reason for rejection of research by medical education journals (Meyer et al., 2018).

Ontology (theory or beliefs about the nature of reality) drives epistemology (theories or beliefs about ways of knowing)(Cohen et al., 2017, Cleland, 2015). These beliefs then form the foundation of the theoretical lens and methodological approach with which the research questions are addressed. Such issues are not generally considered in radiation oncology which usually takes a positivist standpoint; this is understandable due to its foundation in the physical sciences.

Table 4-1 - Contrasting the objectivist and interpretivist scientific paradigms. Adapted from (Cohen et al., 2017, McMillan, 2015 , Varpio et al., 2020).

Paradigm	Objectivist / positivist	Interpretivist / constructivist
<i>Ontology</i>	Reality is objective; knowable & measurable	There are multiple subjective realities; socially constructed
<i>Epistemology</i>	The investigator and participants are independent entities	Perceptions and experiences of researcher and participants are interdependent
<i>Causality</i>	Causality is linear	Causality is multidirectional
<i>Aim</i>	Seeks to explain behaviour and underlying causes	Seeks to understand actions and meanings
<i>Methods</i>	Quantitative (Experiments, surveys, hypothesis testing)	Qualitative (Observation, interviews, participation)

Positivism (Table 4-1), the dominant (and arguably hugely successful) paradigm in the development of the natural sciences, has been challenged, especially in the social sciences, as reductionist (Cohen et al., 2017). In the educational context its limitations are clear - a person's knowledge and experience are not directly measurable, are subjective, and depend on the context and observer (Kettley, 2010).

In medical education a consensus has emerged that these approaches can be complementary rather than compete with each other (Tavakol and Sandars, 2014). As a research paradigm, **Pragmatism** avoids *"the contentious issues of truth and reality [and] accepts, philosophically, that there are singular and multiple realities that are open to empirical enquiry and orients itself toward solving problems in the 'real world'"* (Feilzer, 2009, p.8). Pragmatism accepts that the world can be partly predictable and orderly, and yet at the same time contains uncertainty, complexity, subjectivity and ambiguity (Dewey, 1925, p.47). Rather than enforcing a dichotomy between positivism and constructivism, or quantitative and qualitative methods, pragmatism requires the researcher to use the most appropriate approach to study the phenomenon in question (Onwuegbuzie and Leech, 2005).

My research questions focus on evaluation and design of contouring simulation programmes for assessment and teaching and are rooted in real-world contexts. These requirements fit with **educational design research** - a mixed-methods approach which is outlined below.

4.3 Educational design research

Typically, design researchers want to solve a problem; they see the potential of new technology for teaching ... The type of learning they envision cannot yet be observed in naturalistic settings; hence new settings have to be engineered in which the intended learning processes can be researched and improved" (Bakker, 2019, p.3)

4.3.1 Background

The foundations of educational design research lie in educational psychology (Brown, 1992) & computer science (Collins, 1992). Researchers sought to address questions about the validity of educational theory and impact of educational innovations within their 'real world' contexts rather than a psychology laboratory, and were optimistic that the pace of design innovation in the burgeoning information technology sector could transfer to educational programmes and technologies.

Under a number of different names (including 'design-based research', 'design research', 'design experiments', and 'formative evaluations' (McKenney and Reeves, 2020)), design research in education started to gain traction around the turn of the millennium (The Design-Based Research Collective, 2003), when a seminal series of papers were published in the Journal of the Learning Sciences (Collins et al., 2004). Since then, educational design research has permeated into medical education (Agnew and O'Kane, 2011, Dolmans and Tigelaar, 2012, McKenney and Reeves, 2020).

Despite the variety of names and conceptualisations, there are a number of common characteristics which define the family of approaches (Cobb et al., 2003, Kelly, 2004, Anderson and Shattuck, 2012, Plomp and Nieveen, 2013, Barab, 2014, Bakker, 2019):

- Centred around the design (or re-design) of an educational intervention (which may be technology, learning activities or a whole programme)
- Educational theory is incorporated into the analysis and design stages, and is reflected on during the evaluation
- Iterative process of design, evaluation and re-design (Figure 4-1)
- Conducted in authentic learning environments

- Includes participants in the design process
- Utilises mixed methods

Research quality is not judged using specific educational design research criteria, but using existing quality criteria depending on sub-study methodology - different questions (and theory) are central at different stages of the research programme (McKenney and Reeves, 2020).

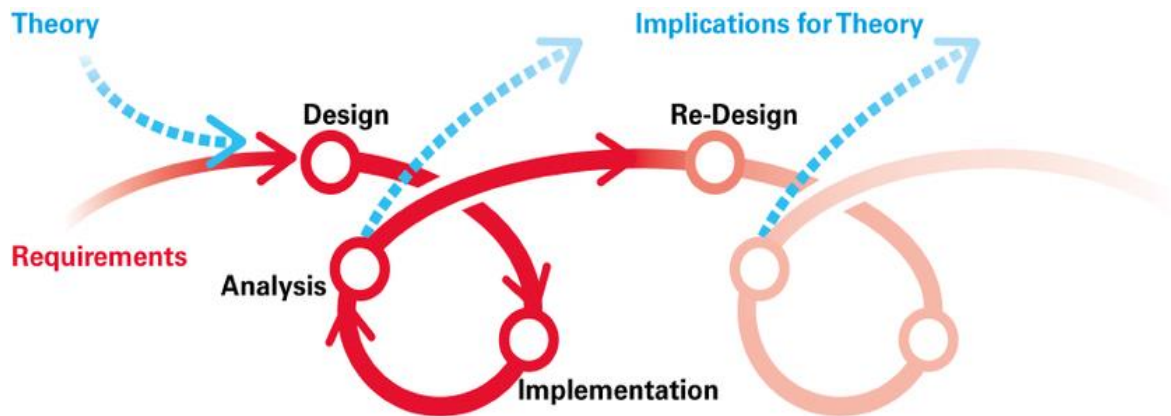


Figure 4-1 - The iterative process of educational design research. Reproduced from Fraefel (Fraefel, 2014, p.9)

Educational design research aims to explore questions about teaching and learning through the design and use of software and/or learning environments (Kelly, 2004), with the ultimate goal of “advancing theory while at the same time directly impacting practice” (Barab, 2014). Educational design researchers often frame these goals within Stokes’ ‘Pasteur’s quadrant’:

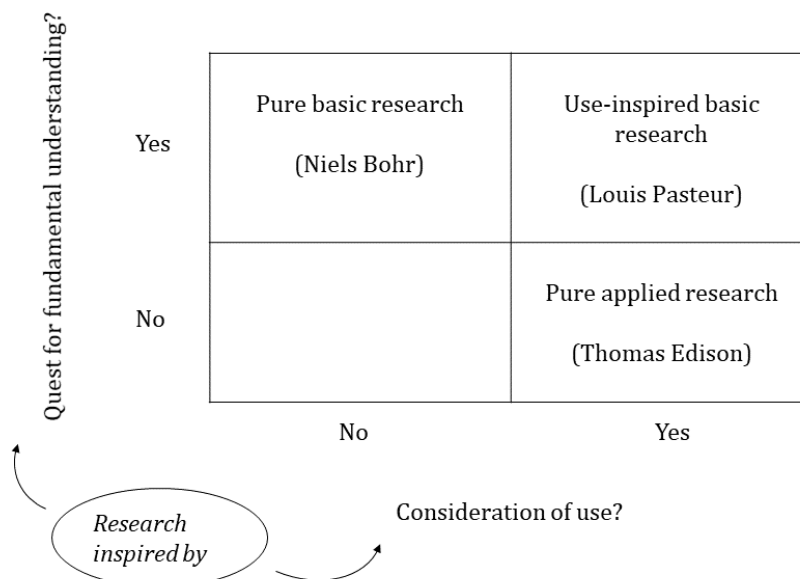


Figure 4-2 - Stokes' classification of research endeavours. Reproduced from Stokes (Stokes, 1997, p.73)

Any initial design and implementation inevitably contains flaws - indeed educational design research has been characterised as “research through mistakes” (Anderson and Shattuck, 2012, p.17), hence the requirement for an iterative approach with a commitment to multiple steps of development and refinement.

4.3.2 Educational design research in health professions education

Educational design research in medical education has recently been reviewed by McKenney & Reeves (McKenney and Reeves, 2020). They explain their generic model for educational design research (Figure 4-3) and stress the dual focus on practice and theory in each of three core phases:

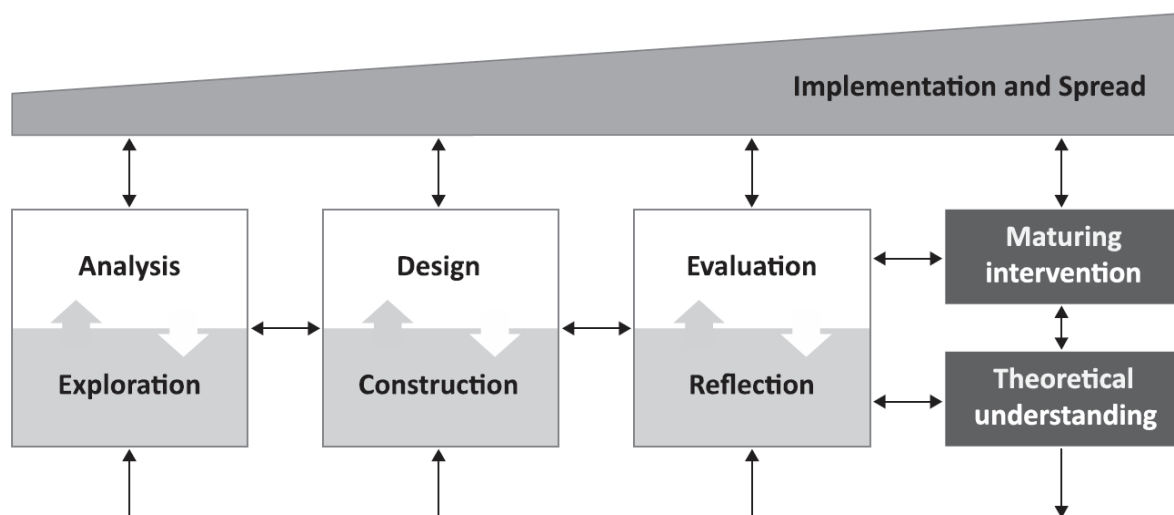


Figure 4-3 - McKenney & Reeves' generic model for educational design research. Mid-level text represents 'practice' and the lower level represents 'theory'. Reproduced from McKenney & Reeves, 2020, p.5

In one example of design research in medical education, Hege et al. (Hege et al., 2017) reported the development of a clinical reasoning tool using virtual patients. The authors first translated a conceptual clinical reasoning framework into software requirements, blueprinting user experience and incorporating adaptive feedback. They then conducted usability pilot testing with cohorts of medical students, although they did not study improvements in students' skills or knowledge at that early stage. In another example, this time relating to practical skills, Ryu et al. conducted an evaluation of three different spinal surgery simulators using a design research framework (Ryu et al., 2017). This allowed the authors to explore the strengths and limitations of each simulator for different groups of learners, and propose unique scenarios and cohorts for each to augment learning to be tested with future experimental designs. This example of researchers asking “what works, how and for whom?” rather than proceeding straight to the

empirical “does it work?”, as highlighted in Chapter 3 (see Section 3.4.2), may have been aided by an educational design research framework.

There is considerable overlap between educational design research and action research (Anderson and Shattuck, 2012) - both are rooted in pragmatism and constitute applied research interested in addressing problems within their real-world context (Cole et al., 2005). However in action research design is optional and there is less focus on the incorporation, testing & refining of theory (Barab and Squire, 2004). In addition, in educational design research the researcher is able to set their own goals and methods for the intervention design, rather than facilitating the participants’ design processes (Wang and Hannafin, 2005). Therefore, this programme of research is better categorised as educational design research.

4.3.3 Cautions & validity threats

Educational design research opens up the tantalising possibility of pairing innovation with the advancement of theory, but this newly-conceived approach has several limitations which need to be borne in mind by the researcher.

The researcher ‘as designer’ (or at least as part of the design team) has a vested interest in the ‘success’ of their design project. This may lead to the evaluation of one design rather than comparing alternate designs (Kelly, 2004), as in Hege et al.’s study above, but not for Ryu et al. who evaluated three different designs.

Causal claims from observational research are to be interpreted with caution due to multiple uncontrolled variables. As Barab puts it: *“Critics of DBR [design-based research] tend to be advocates of controlled experimental methodologies. These critics argue that DBR does not provide empirical evidence to ground claims; at best it can provide formative insights that must then be tested through more controlled experimentation”* (Barab, 2014, p.152). Generalisability from these ‘real-world’ studies is limited by their unique context, but that is not to say that lessons cannot be learned across contexts - indeed the triangulation of findings such as ‘design principles’ (Van den Akker, 1999) that are replicated across contexts can provide validity evidence.

Educational design research requires methodological flexibility - therefore researchers must have a good understanding of the merits of possible approaches and select these based on the study purpose (McKenney and Reeves, 2020). This could be a challenge within a programme of doctoral research, given (by definition) the limited experience of the researcher. Exploratory studies with small sample sizes are especially vulnerable to their results being skewed by chance

events, but larger samples can limit the researcher's ability to conduct in-depth qualitative analysis (Sandelowski, 1995).

The iterative nature of educational design research leads to long-term commitment & associated resource implications (Anderson and Shattuck, 2012). The initial iteration may not produce meaningful outputs - instead a sustained programme of research may be required to produce these - but this is in keeping with much of educational research (McGaghie et al., 2014).

The novelty of the approach encapsulated by educational design research means that there has not been sufficient time to evaluate whether it has delivered on its original promise, but exemplar research programmes (for example: Vesper, 2014) show that it is an approach that can produce effective educational innovation which can then be applied to new contexts.

4.4 Programme of research & study cohorts

Chapters 2 and 3 highlight the importance of considering assessment and teaching of radiotherapy contouring both in groups of trainees and accredited practitioners, as their cognitive schemata, practices, and responses to interventions may differ.

My involvement in contouring education and quality assurance in the EMBRACE-II trial in locally advanced cervix cancer was a unique opportunity to explore these domains in a group of 'expert' practicing clinicians. As discussed in Chapter 2, assessment & education is already established for accredited practitioners within the clinical trial setting in order that trial results are not confounded by sub-optimal radiotherapy. The middle part of this thesis, as part of the 'analysis and exploration' phase of educational design research, contains two case studies of radiotherapy contouring quality assurance within EMBRACE-II which highlight the issues with radiotherapy assessment and education in this context.

For UK trainees, contouring education is mainly carried out via working through cases at their local radiotherapy centre (Evans et al., 2019a). As yet, to my knowledge, there is no established national or regional programme of contouring education or assessment, although a group from Cardiff in the UK have embarked on such a programme (Evans et al., 2019b), and the Royal College of Radiologists have developed self-directed 'anatomy for radiotherapy' modules. Therefore to conduct contouring education research with groups of trainees I worked within existing training programmes which generally revolve around regular local or regional sessions which rotate through different topics aiming to give overall coverage of the oncology curriculum. I was able to recruit groups of trainees locally (Chapter 8), regionally (Chapter 9) and internationally (Chapter 9).

The EMBRACE-II clinical trial is outlined below, after which I go on to outline the sub-studies and cohorts investigated in this programme of doctoral research.

4.4.1 The EMBRACE-II trial

The background to the EMBRACE-II trial was explained in Chapter 1 (see Section 1.3.3). As mentioned there, the EMBRACE trial group has been a key force in validating and driving forward the science and practical implementation of image-guided brachytherapy (IGABT) for cervical cancer (Pötter et al., 2018). The Retro-EMBRACE and EMBRACE-I trials have also benchmarked world-leading outcomes for cervical cancer patients in terms of local and regional control (Sturdza et al., 2016, Tan et al., 2019a, Pötter et al., 2020). It is therefore an interesting cohort to study contouring practice and variation as one can relate the documented contouring variation within the EMBRACE-I group to excellent clinical outcomes - as seen in Chapter 2 this link is not always present.

Clinicians from the EMBRACE trial group are also extremely active in education for gynaecological radiotherapy, particularly for IGABT. They form the core faculty for the 'ESTRO image-guided radiotherapy and chemotherapy in gynaecological cancer' teaching course which has been attended by over 2500 participants over the last 17 years (Tan et al., 2020).

The EMBRACE-II trial (NCT03617133, www.embracestudy.dk) is an international prospective single cohort interventional study of IMRT and MRI-based IGABT in loco-regionally advanced cervix cancer, and follows on from Retro-EMBRACE and EMBRACE-I. It aims to recruit over 1400 patients over the course of 2016-2021, and as of the end of 2020 had recruited over 1000. The aims of the study are listed in Table 4-2:

Table 4-2 - Aims of the EMBRACE-II study. Reproduced from the EMBRACE-II protocol (Potter, 2016)

- | |
|--|
| <ul style="list-style-type: none"> • To systematically apply IMRT with daily IGRT as well as advanced image guided adaptive BT in a prospective multi-centre setting • To systematically implement a dose prescription protocol for IGABT • To implement systematic contouring, prescription and reporting for EBRT CTV and OaRs. • To administer EBRT in different targets which are adapted to the risk of nodal and systemic failure: to improve para-aortic and systemic control in high-risk patients and not to decrease lymph node control in low risk and intermediate risk patients • To systematically administer simultaneous chemotherapy to EBRT to reach prescribed dose in as many patients as possible, in particular in high risk patients |
|--|

- To benchmark an outstanding high level of local, nodal and systemic control as well as survival with application of advanced EBRT, BT and chemotherapy within limited overall treatment time
- To benchmark a low incidence of intermediate and major morbidity as well as a high level of quality of life with application of advanced EBRT, BT and chemotherapy

The main differences in radiotherapy technique when compared with EMBRACE-I are:

- Systematic application of brachytherapy prescription and dose-volume targets, with dose de-escalation for small tumours, reduction of vaginal brachytherapy source loading, and dose escalation for large tumours mediated by an increase in interstitial needle application
- Obligatory use of IMRT at a dose of 45Gy in 25 fractions for the EBRT component of treatment, with an 'internal target volume' concept (the first in any large-scale clinical trial in cervical cancer) to account for target motion, with daily image-guided radiotherapy
- Adaption of the IMRT elective lymph node clinical target volume to the patient's risk of lymph node metastases
- Standard practices (dose and target definition) for lymph node boosting for IMRT, using probabilistic target coverage

Consistent implementation of the radiotherapy contouring protocol for both EBRT and brachytherapy is important to achieving the aims of this trial, both in terms of evaluating the effect of the above interventions on local and regional control, survival and toxicity, and in validating radiation dose-response relationships.

4.4.2 Structure of studies within the thesis

The phases of educational design research (analysis & exploration, design & construction, evaluation & reflection) form a structure for this thesis. Chapters 2 & 3, and 5 & 6 represent the analysis and exploration phase. Chapters 5 and 6 are real-world case studies of radiotherapy contouring assessment and teaching embedded in the EMBRACE-II trial quality assurance process, which was designed prior to this doctoral research. These data highlight some of the issues with assessment of radiotherapy contouring as it is currently conducted, and explore the impact of an online education programme to supplement quality assurance.

Chapter 7 represents the design & construction phase, where I present my case for the development of a new low-fidelity contouring simulation based on principles from educational theory and best practice, and describe the initial development and early testing.

Chapters 8 and 9 represent the evaluation and reflection phase, where I report a detailed usability study of this simulation (Chapter 8) and three pilots using the simulation to assess and teach cervical cancer contouring in groups of UK-based and international trainees and EMBRACE group clinicians (Chapter 9). Chapter 10 (conclusion, and the start of the next design research iteration) brings together findings from this programme of research together with some reflections and possibilities for future avenues of research.

5 EMBRACE-II EBRT accreditation: online education to support radiotherapy quality assurance

5.1 Introduction

As explained in Chapter 2, robust radiotherapy quality assurance (RTQA) is required to minimise contouring variation and other protocol deviations which may affect treatment outcome.

Before being allowed to recruit patients in EMBRACE-II, centres were required to undergo a comprehensive RTQA evaluation comprising a compliance questionnaire, IMRT contouring, and IMRT dose planning. In addition, centres that did not participate in the first EMBRACE study (EMBRACE-I) needed to be assessed on brachytherapy contouring and on prospective data from 5 real patients treated according to the EMBRACE-II protocol.

In EMBRACE-I, the radiotherapy contouring and dose planning RTQA programme involved 30 centres submitting two of their own cases for central review (Kirisits et al., 2015). While the programme was successful in identifying and correcting common protocol deviations at an early stage, the process was time-consuming, often requiring repeated communication between the study office and centre personnel to resolve digital data transfer issues. This was considered impractical for EMBRACE-II due to the larger number of participating centres and limited resources. The RTQA process often involves centres downloading DICOM datasets of benchmark cases into their clinical software for contouring and dose planning and re-uploading the results for central review. While several platforms have been developed to facilitate this process, the challenges of such digital exchange remain considerable with significant cost and manpower implications (Weber et al., 2011, Bekelman et al., 2012).

A different RTQA process was therefore developed for EMBRACE-II. Benchmark cases (Gwynne et al., 2013) for contouring and dose planning were hosted via a Moodle open-source learning management system accessed through a website (Cambridge Cancer Medicine Online; <https://ccmo.co.uk>). For contouring, participants were required to use a high-fidelity online contouring tool (the Addenbrooke's Contouring Tool - Figure 5-1) which eliminated the need for digital transfer.



Figure 5-1 - The Addenbrooke's Contouring Tool (ACT)

The use of a learning management system provided opportunity to develop an online continuous medical education (CME) programme for all study participants to highlight and reinforce key aspects of the protocol. The focus of this programme was on the IMRT component of the protocol as it involves several interventions which imply a change of practice for many centres, including daily image guidance, an individualised internal target volume for the primary tumour (ITV-T), an elective lymph node clinical target volume (CTV-E) adapted to the patient's risk of lymph node metastases (Figure 5-2, Table 5-1).

Table 5-1- Risk groups for defining the elective lymph node clinical target volume in EMBRACE-II. Reproduced from the EMBRACE-II protocol (Pötter et al., 2016a).

Lymph node risk group	Criteria
Low risk "Small pelvis"	Tumour size $\leq 4\text{cm}$ AND stage IA/IB1/IIA1 AND squamous cell carcinoma AND no uterine invasion
Intermediate risk "Large pelvis"	Not low risk No high risk features
High risk "Large pelvis + para-aortic region"	≥ 1 pathologic lymph node at common iliac or above OR ≥ 3 pathologic lymph nodes

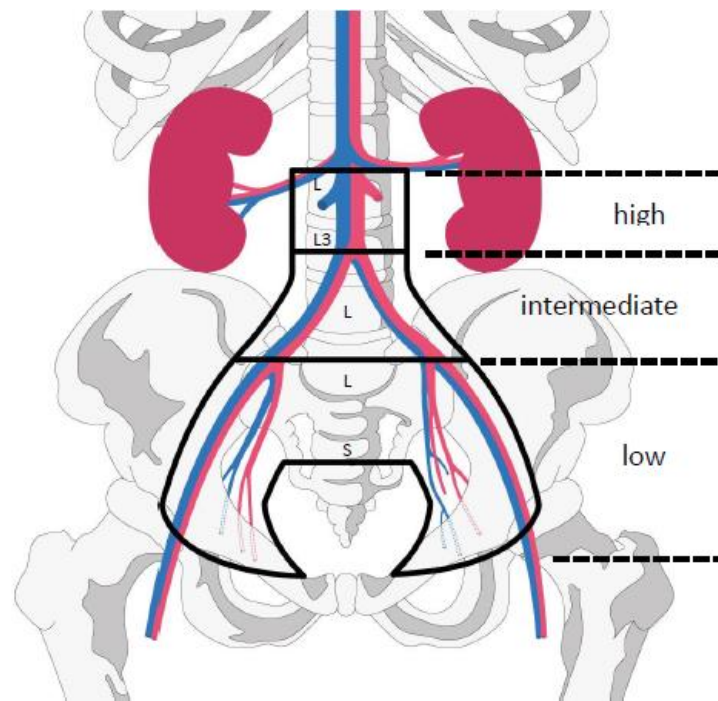


Figure 5-2 - Risk-adaptive elective lymph node CTV (CTV-E) in EMBRACE-II protocol

The CME programme also included practice cases for online contouring, which as explained in Chapter 2 (Section 2.4.4), has been shown to be helpful in improving contouring homogeneity and adherence to contouring guidelines.

The results of the RTQA programmes for IMRT dose planning (Seppenwoolde et al., 2019) and IGABT contouring (Chapter 6) are reported elsewhere. This chapter reports the results of the EMBRACE-II IMRT contouring RTQA programme and analyses user engagement with the supporting CME programme.

The aims of this chapter are to:

- Present the overall performance of clinicians applying new external beam concepts in cervical cancer
- Identify consistently repeated errors in contouring
- Explore the uptake and impact of the online educational programme
- Analyse the relationship between geometric overlap and manually assessed clinical acceptability (score)

My contribution

This study is the result of collaborative research within the EMBRACE-II trial group. Collaborators and their affiliations are listed below in Appendix Table A.5-1.

The EMBRACE-II trial management group designed and initiated the EMBRACE-II RTQA process in 2016. I was involved in the assessment of contouring submissions during the latter part of the accreditation. I am grateful for the assistance of Dr Hatem Helal in coding an initial MATLAB import of the contour data from the Addenbrooke's Contouring Tool.

For this study I collated the data, conducted the data analysis and wrote the first manuscript draft. This was then revised and agreed by all collaborators before being submitted for publication in August 2019. We received a response in December 2020: major revisions were requested. I revised the manuscript and drafted responses, and re-submitted the manuscript in February 2020 which was accepted immediately. This chapter is therefore based on published work:

- Implementing an online radiotherapy quality assurance programme with supporting continuous medical education – report from the EMBRACE-II evaluation of cervix cancer IMRT contouring. Duke SL, Tan LT, Jensen NBK, Rumpold T, de Leeuw A, Kirisits C, et al. *Radiotherapy and Oncology* 2020 Jul; 147: 22-29.

5.2 Methods and materials

The EMBRACE-II RTQA and CME programmes were established in early 2016 and have been available to all study participants since then.

5.2.1 RTQA programme

Accreditation case

For IMRT contouring accreditation, the principal investigator (PI) from each centre was required to contour on one benchmark case. Non-PIs were also encouraged to submit contours and receive formative feedback. The accreditation case was a stage T2bN1M0 cervical squamous cell carcinoma with 3 pathologic lymph nodes (defined as high-risk in the EMBRACE-II protocol - see Figure 5-2 and Table 5-1) which meant that clinicians should extend the elective lymph node target volume to include the para-aortic lymph nodes.

Clinicians were required to contour on axial T2-weighted MRI images, obtained in the treatment position, fused to planning computed tomography (CT) images. Additional case information (clinical history + examination findings, diagnostic multi-planar MRI + positron emission tomography (PET-CT) images and reports) were embedded within the contouring tool.

The regions of interest (ROIs) assessed were defined in the protocol, and were adapted from the ICRU89/GEC-ESTRO recommendations (Haie-Meder et al., 2005):

- **GTV-T_{init} = initial gross tumour volume.** This is assessed at diagnosis using a combination of clinical examination findings and imaging.
- **CTV-T_{HRinit} = initial high-risk clinical target volume.** This is the region at highest risk of recurrence after treatment. For cervical cancer, this comprises the GTV-T_{init} and any uninvolved cervix.
- **CTV-T_{LRinit} = initial low-risk clinical target volume.** This represents regions (which are anatomic compartments) at risk for potential microscopic spread from the primary tumour. In locally advanced cervical cancer this comprises the whole parametria, the whole uterus, and the vagina extending 2cm caudally from the **GTV-T_{init}**.
- **CTV-N = pathological lymph nodes clinical target volume(s)** (i.e. involved with tumour). This volume covers the gross lymph node tumour (GTV-N) as seen on both MRI and CT with an additional 0-3mm margin.
- **CTV-E = elective lymph node clinical target volume.** This volume represents areas of potential microscopic lymph node spread and also includes pathological lymph nodes if present. The accreditation case was classified as “high risk” (Table 5-1, Figure 5-2) i.e. para-aortic region should be treated.
- **ITV-T = internal target volume (local tumour).** This begins with the CTV-T_{LRinit} expanded geometrically (10mm cranio-caudal + antero-posterior, 5mm laterally) which is then individually edited depending on the position of the target relative to organs at risk, and predicted patterns of motion.

Detailed descriptions of these ROIs are given in the EMBRACE-II protocol (Pötter et al., 2016b) and elsewhere (Pötter et al., 2016a, Pötter et al., 2018).

Scoring system

The first cohort of 20 submissions was assessed jointly by five oncologists (IJ, NJ, JL, RN, LT - see Appendix Table A.5-1) at a face-to-face meeting in May 2016. Each ROI was compared to a consensus reference and was scored as 2 (excellent), 1 (fair) or 0 (revision required). At this meeting the principles for assigning the scores were decided – scores of 0 were assigned to errors demonstrating a fundamental flaw in conceptual understanding and/or potential clinically significant consequences on loco-regional control (see Figure 5-5 for example scoring).

Subsequent submissions from September 2016 were assessed remotely by two oncologists using the previously agreed principles, however there were no written criteria for assigning scores. Any disagreements or queries from the assessors were discussed with the wider group to finalise a consensus score. The organs at risk were also reviewed and qualitative comments made as necessary, but they were not scored.

Clinicians were given individualised feedback on their scores after each submission. For contours receiving a 0 or 1 score, additional qualitative comments were provided. A total score of ≥ 9 out of 12, with no ROIs requiring revision (i.e. scoring 0/2), was needed to pass. Clinicians who did not pass were required to revise their contours on the same case and resubmit them for re-evaluation.

5.2.2 Online CME programme

Initial CME content included an IMRT contouring atlas, practice cases for contouring and dose planning, and quizzes. Optional feedback links were provided to gather participants' opinions on each resource. Learning analytic data was collected by the learning management system each time a participant accessed a page or resource in the CME or accreditation programmes. The CME programme was available to all staff groups but this analysis is restricted to oncologists.

Pre-accreditation Questionnaire

An online questionnaire was placed at the start of the CME programme material. This collected information about the clinicians' experience, previous training, guideline use, and aspects of cervix cancer IMRT contouring that they found difficult.

Delineation practice cases

Two practice delineation cases were provided - a stage T2bN1M0 patient (intermediate risk - therefore treated with a standard pelvic lymph node volume [Figure 5-2]) and a stage T1bN1M0 patient (high-risk). Clinical information and diagnostic information was embedded in the ACT delineation tool. A reference contour was provided by a single expert for each case. For each of these cases participants could practice delineation on any of the EBRT ROIs. Once they submitted their contours, a reference contour was visible for comparison.

Quizzes

Quizzes were designed as quick ways to reinforce aspects of the protocol and its interpretation. Two were created, with 10 questions each on the following topics:

- General EBRT concepts (Quiz 1)
- EBRT image guidance and planning (Quiz 2)

Other resources

A contouring atlas was included in the materials along with a 'quick contouring guide'. An ITV-T step-by-step guide (in presentation format) was added in September 2016 in response to participant difficulties.

5.2.3 Data Analysis

A retrospective analysis of first-attempt submissions was carried out (SD). Qualitative feedback comments for ROIs with scores of 0 or 1 were grouped into themes to identify common errors. As there are reports of marginal recurrences when IMRT was initially implemented in other tumour sites (Eisbruch et al., 2004, Schoenfeld et al., 2008, Chen et al., 2011), the potential implications of under-contouring errors on loco-regional control were assessed (IJ, LT). As there is no compelling evidence linking over-contouring in IMRT with excess toxicity, the implications of over-contouring errors were not assessed.

The Jaccard conformity index (JCI; see Figure 5-3), a measure of geometric overlap (Hanna et al., 2010), for each participant VOI was calculated using MATLAB (The Mathworks Inc., 2018) and was compared to the expert-assigned score.

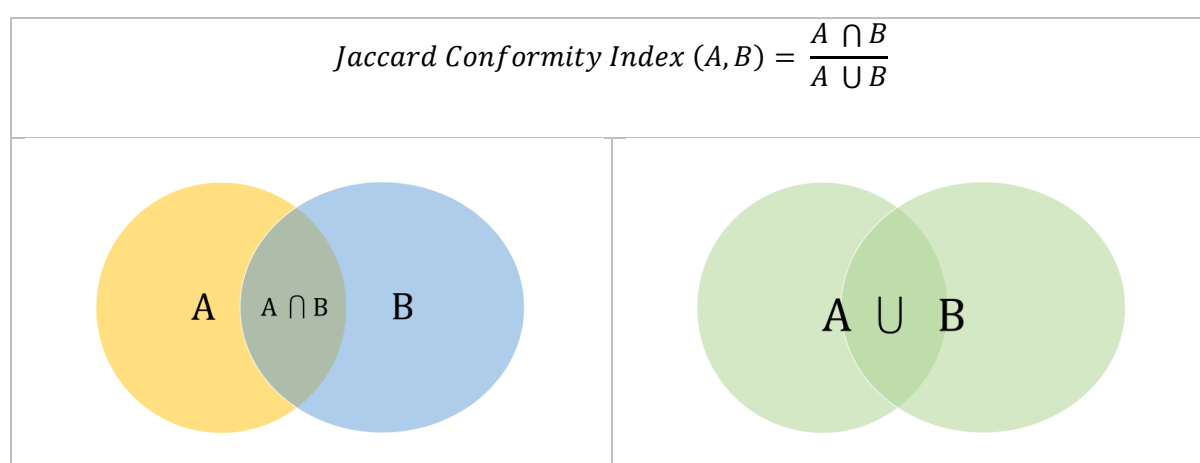


Figure 5-3 - Derivation of the Jaccard conformity index

Learning analytic data, collected by the learning management system each time a participant accessed a resource, were analysed.

Pass rates and scores for regions of interest were compared between groups using the chi-squared test and two-sample t-test respectively.

5.2.4 Ethical approval

The EMBRACE-II trial has ethical approval and is sponsored by the Medical University of Vienna. Consent for analysis of participant data for the purposes of education and research was obtained at the point of entry to the delineation tool. All centres actively recruiting patients also have national and local ethical approval.

5.3 Results

5.3.1 Accreditation case

78 clinicians (including 9 non-PIs) from 67 centres in 24 countries participated in the IMRT contouring evaluation. Assessments began in May 2016 and were completed by July 2018.

32 clinicians (41%) passed at the first attempt. 40 revised and re-submitted their contours following individualised feedback and of these, 34 (85%) passed on the second attempt. This led to an overall pass rate of 85% after the second attempt. A further 5 passed on their third attempt.

Of the 9 non-PIs who submitted to the accreditation case; 6 failed at the first attempt, and in all but one of these circumstances the PI at their centre had passed first time.

Figure 5-4 shows the mean score for each ROI at the first contouring attempt. The ROIs that received the lowest scores on the first attempt were CTV-E (average 1.01), ITV-T (1.06) and CTV-T_LR_{init} (1.22). Similarly, the ROIs most commonly scored as “0” / “requiring revision” were the CTV-E (25/78 = 32%), ITV-T (23/78 = 29%) and the CTV-T_LR_{init} (11/78 = 14%).

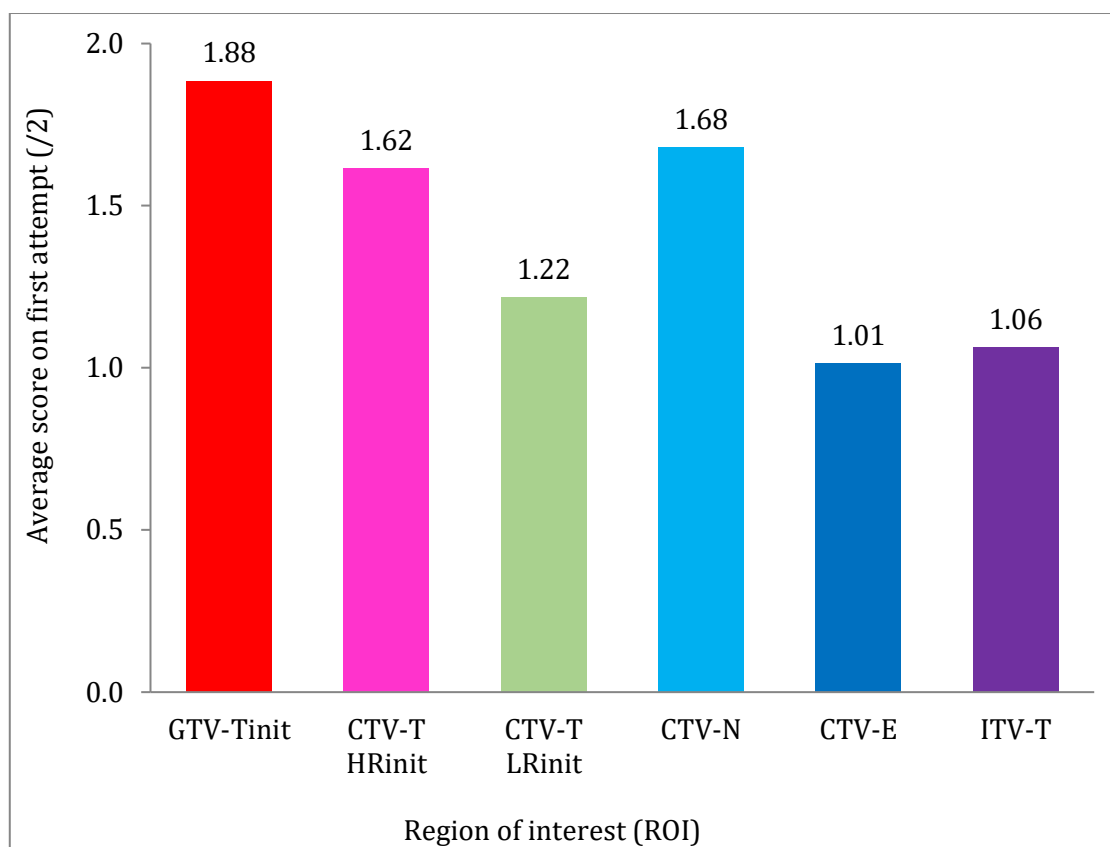


Figure 5-4 - Average scores per ROI for EMBRACE-II EBRT delineation accreditation case

Analysis of Qualitative Errors

Overall, 60 types of error across the six ROIs were identified from the qualitative feedback comments. The most common and/or clinically significant errors are shown in Table 5-2; 12 involved under-contouring while 10 were over-contouring. Five of the under-contouring errors were assessed as having high implications for loco-regional control as the errors involved regions of high dose gradients and/or gross geographical miss. Another four under-contouring errors were estimated to have moderate implications for loco-regional control. 30/60 errors related to only 1 or 2 clinician contours (i.e. rare), whereas only 4 errors were made by more than 16 clinicians (i.e. more than 20% of the group).

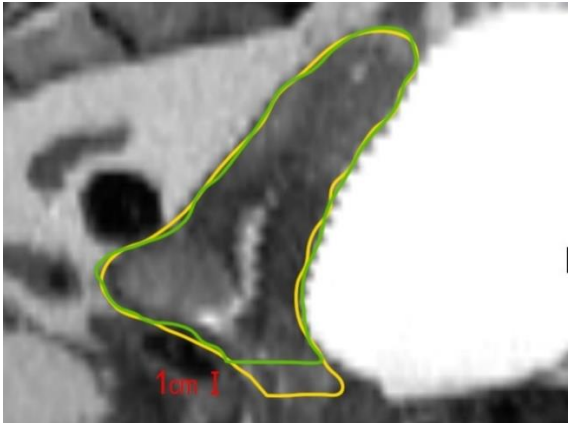
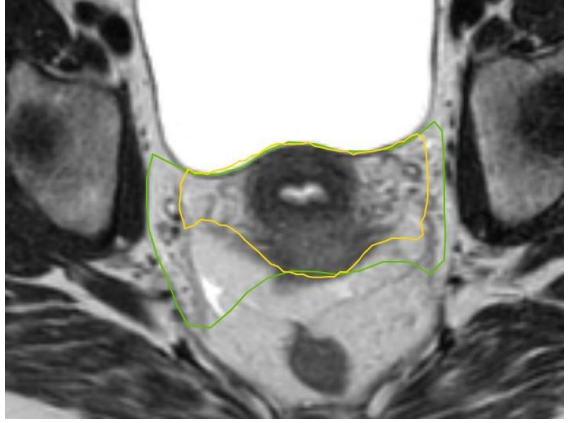

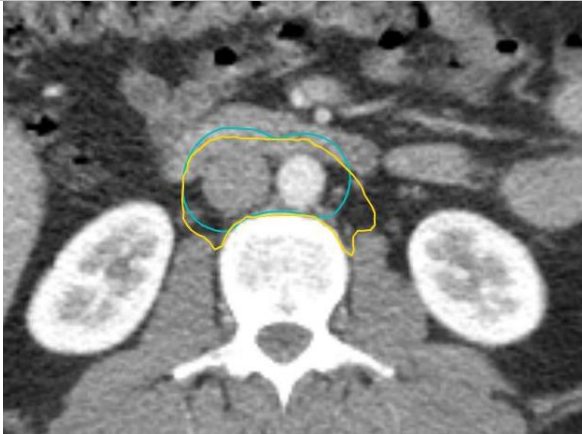
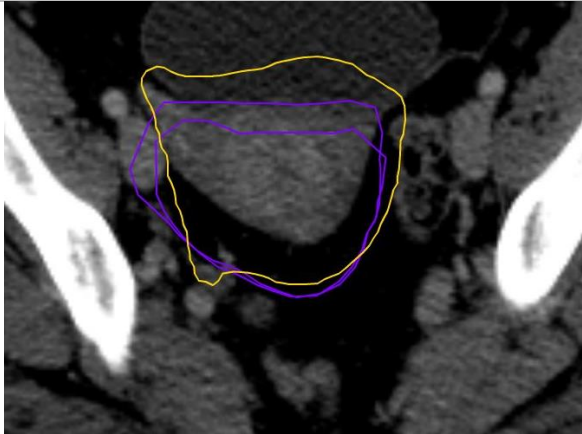
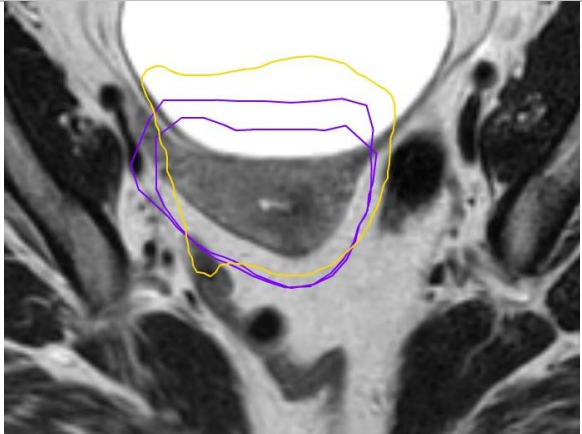
<p>A.</p> 	<p>B.</p> 
<p>A & B: CTV-T_LR for two participants (green). A: The contour is too superior along the vaginal axis. In this case, the JCI was 0.85 despite the high risk of geographic miss (scored '0'). B: The participant contour extends laterally outside the parametrium into the elective nodal volume (scored '1').</p>	
<p>C.</p> 	<p>D.</p> 
<p>C&D: CTV-E for 2 participants (blue). C: The participant has not contoured the para-aortic nodes (scored '0'). D: The participant has missed the left lateral para-aortic nodes (scored '1').</p>	
<p>E.</p> 	<p>F.</p> 
<p>E & F: ITV-T_LR for two participant contours (purple). On MRI (left image), the contours appear satisfactory but on CT (right image), one participant did not cover the uterus anteriorly (scored '0') while the other just covered the uterus but margin was insufficient for movement during treatment (scored '1').</p>	

Figure 5-5 - Examples of participant errors and assigned scores for various regions of interest. Consensus contours are in yellow

Table 5-2 - Common and/or clinically significant errors seen in first submissions for EMBRACE-II IMRT benchmark case

VOI	Description	Participants	Type	Implications for loco-regional control	Comment
GTV-T_{init}	A variety of errors at low frequency				
CTV-T_{HRinit}	Anterior lip of cervix missed	23%	Under-contouring	Moderate/ High	Impact depends on length of vagina contoured
CTV-T_{LRinit}	Paravaginal tissue not included	33%	Under-contouring	Low	Covered by ITV-T expansion
	Uninvolved vagina too short	9%	Under-contouring	High	High dose gradient at edge of field
	Uninvolved vagina too long	13%	Over-contouring		
	Parametrial border too narrow	3%	Under-contouring	Low	Merges into CTV-E
	Parametrial border too wide	13%	Over-contouring		
	Mesorectum/rectum included	14%	Over-contouring		
GTV-N/CTV-N	Pathological node completely/partially missed	13%	Under-contouring	Uncertain	Value of nodal boost being assessed
CTV-E					
Paraaortic nodes	Not contoured	13%	Under-contouring	High	Gross geographical miss
	Contoured but superior border too low	9%	Under-contouring	High	High dose gradient at edge of field
	Contoured but superior border too high	19%	Over-contouring		
	Left lateral space missed	31%	Under-contouring	High	High dose gradient at edge of field
Pre-sacral nodes	Inferior border too high	9%	Under-contouring	Moderate	Uncommon area for recurrence
	Inferior border too low	9%	Over-contouring		
External iliac nodes	Contours extended into inguinal region	9%	Over-contouring		

VOI	Description	Participants	Type	Implications for loco-regional control	Comment
Obturator nodes (CTV-E)	Inferior border too high	17%	Under- contouring	Moderate	Merges into CTV-T_LR _{init}
	Inferior border too low	6%	Over-contouring		
ITV-T	Anterior border too tight	32%	Under- contouring	Moderate	Uncommon area for recurrence
	Anterior border too generous	4%	Over-contouring		
	Posterior border too tight	15%	Under- contouring	High	High dose gradient at edge of field
	Posterior border too generous	3%	Over-contouring		
	Additional margin added inferiorly to vagina	13%	Over-contouring		

5.3.2 Conformity Index Analysis

56/78 (72%) first-attempt contours were available for JCI analysis - the remaining 22 first-attempt contours had been over-written by subsequent submissions due to the initial database design (which was subsequently modified).

There were discrepancies between pass rates using various JCI cut-offs compared with expert assessments (Table 5-3). Using a JCI cut-off of 0.7, a commonly used threshold for clinically adequate delineation in studies (Fokas et al., 2015), only 45% (157/342) of all contours passed by the experts would have passed ("true-positive") while 55% (185/342) would have failed ("false-negative"). Moreover, 13% (6/46) of the contours that failed expert assessment would also have passed ("false-positive"). Using a more lenient cut-off of 0.65, the true-positive rate was 60% (205/342) but the false-positive rate was 37% (17/46). At a higher JCI cut-off of 0.75, all but one of the failing contours would have been identified but the false-negative rate was 74% (252/342).

Table 5-3 - Comparison of pass rates for 56 participants using various JCI cut-offs with expert assessments

VOIs	JCI cut-off	Expert-assessed as "Pass" (true-positive)		Expert-assessed as "Fail" (false-positive)	
Overall	n	342		46	
	0.65	205	(60%)	17	(37%)
	0.7	157	(45%)	6	(13%)
	0.75	90	(26%)	1	(2%)
GTV-T _{init}	n	56		0	
	0.65	12	(21%)		
	0.7	6	(11%)		
	0.75	1	(2%)		
CTV-T _{HRinit}	n	53		3	
	0.65	53	(100%)	1	(33%)
	0.7	49	(92%)	1	(33%)
	0.75	36	(68%)	0	
CTV-T _{LRinit}	n	47		9	
	0.65	43	(91%)	1	(11%)
	0.7	40	(85%)	1	(11%)
	0.75	28	(60%)	1	(11%)
CTV-N*	n	51		3	
	0.65	4	(8%)	0	
	0.7	0		0	
	0.75	0		0	
CTV-E	n	39		17	

	0.65	36	(92%)	10	(59%)
	0.7	25	(64%)	2	(12%)
	0.75	12	(31%)	0	
ITV-T_LR	n	44		12	
	0.65	38	(86%)	5	(45%)
	0.7	24	(55%)	2	(18%)
	0.75	12	(27%)	0	

5.3.3 CME Programme

Pre-accreditation Questionnaire

39 out of 78 clinicians (50%) who submitted contours for the accreditation case responded to the pre-accreditation questionnaire. Nearly all of the clinicians who responded (36/39 = 92%) had contoured independently on more than 10 patients undergoing IMRT for cervix cancer. All were using IMRT routinely in clinical practice, but only 15 (38%) routinely delineated a formal ITV-T. 31/39 (79%) clinicians had experience of delivering pelvic IMRT in non-gynaecological tumour types such as anal, rectal, prostate and bladder cancers.

19 clinicians (49%) had participated in formal training for cervix cancer IMRT delineation – this was most commonly in the form of a live teaching course (13 out of these 19 = 68%), but also included e-learning (5/19 = 26%) and trial quality assurance (5/19 = 26%). Of the remaining clinicians, 8/39 (20%) had visited another department but not participated in a formal training course. 7/39 (18%) reported that their only training was from another clinician in the same department, and 5/39 clinicians (13%) had not received any training.

The most commonly reported areas of contouring difficulty were the ITV (14/39 = 36%), vagina (12 = 31%), and parametrium (12 = 31%). Only 5 (13%) of clinicians reported difficulty with the CTV-E. 10 (26%) clinicians reported ‘no particular difficulty’ with contouring, although 8 of these 10 required revisions to their first submission for the accreditation case.

The most commonly used guidelines for cervix cancer IMRT delineation at the time of questionnaire completion were the international consensus guidelines published by Lim et al. (Lim et al., 2011) with 31/39 clinicians (79%) having used these.

CME contouring

58 / 78 (74 %) of the clinicians who submitted the accreditation case accessed at least one of the practice cases (Figure 5-6), and of these 29 / 58 (50%) saved contours. In addition to the 78 clinicians above, a further 9 clinicians (all of them non-PIs) saved contours on the practice cases but chose not to submit contouring for the accreditation case.

There was no statistically significant difference between the pass rates of clinicians who accessed CME contouring prior to their first submission versus those who did not (63% versus 50%, $p=0.28$), or their total scores (Average first attempt score 8.9 versus 8.1, $p=0.12$).

CME Quizzes

21/ 78 clinicians (27 %) accessed the quizzes, and of these 13 / 21 (62 %) completed them. The median time to complete a quiz was 10 minutes. The average score in quiz 1 was 70%, with only 1 user getting all answers correct on the first attempt. The quiz question receiving the lowest score (42 %) was on the CTV-E, which was also the lowest scoring ROI in the accreditation exercise.

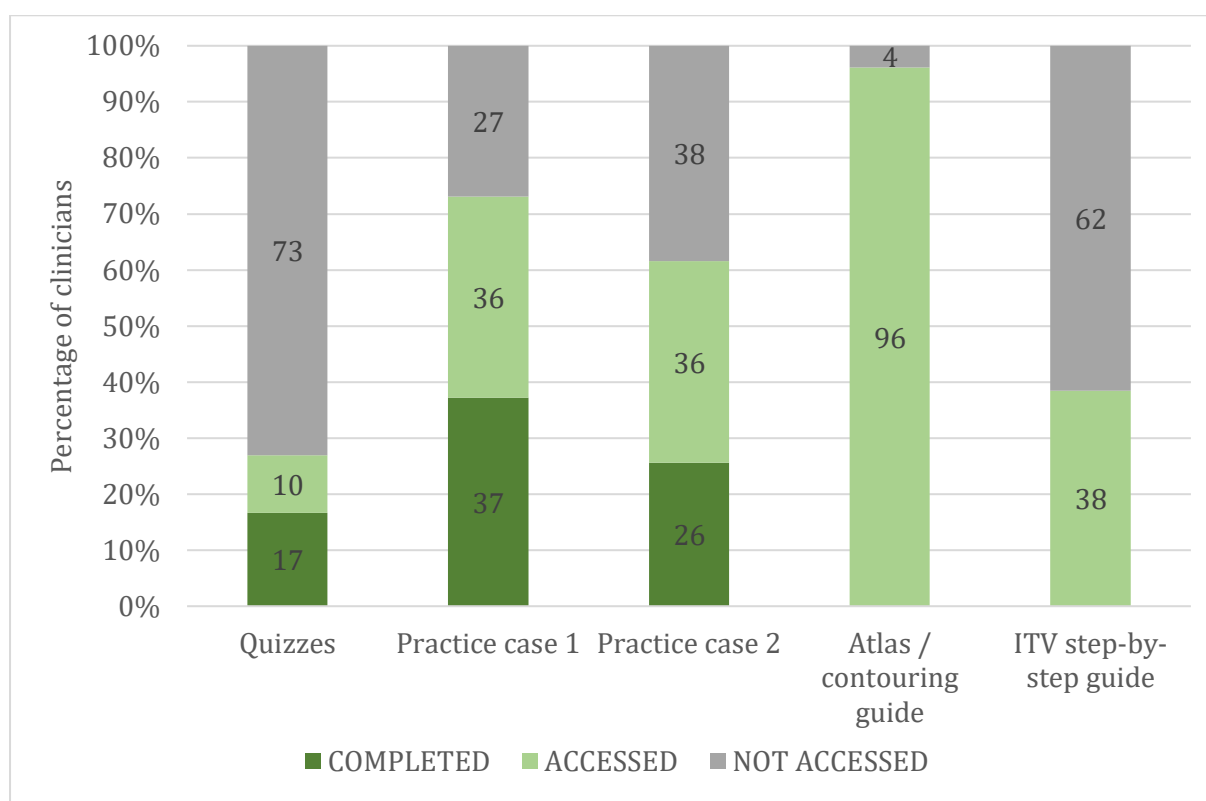


Figure 5-6 - Engagement of EMBRACE-II clinicians with optional CME content

Other CME content

74 (96%) clinicians accessed the contouring atlas and quick contouring guide, often multiple times. This was often the first resource that was accessed. Very few clinicians left comments in the optional 'feedback' sections of the course, for any learning resource.

5.4 Discussion

This study has shown the feasibility of an online RTQA programme for IMRT contouring that requires minimal digital data exchange. 78 clinicians from 67 centres in 24 countries were assessed (in addition to dose planning) despite limited resources. In the setting of an international trial spanning more than 60 centres, 3 continents and 12 time zones, the accessibility, cost and convenience of the online format was a significant boon.

This is the first study to couple a comprehensive online CME programme, including quizzes and self-directed learning materials, with summative assessment of delineation in a benchmark accreditation case. In terms of number of participants, this is the largest study of a benchmark EBRT delineation case in cervix cancer to date. It is also the first study to report on implementation of the ITV concept in a large-scale cervix cancer trial.

Performance across ROIs

The VOIs that received the lowest scores were the CTV-E, ITV-T and CTV-T_{LR_{init}}. Difficulty with the ITV-T was anticipated as it was unfamiliar - in the pre-accreditation questionnaire, only 38% respondents were routinely delineating the ITV-T prior to the trial. In addition, the ITV-T is a complex concept requiring the clinician to predict the movement of the cervix and uterus during treatment, based on one or more pre-treatment images, and adjust the margins added accordingly (Tan et al., 2019b). For example, if the rectum is relatively full, the posterior margin of the ITV-T should be increased to minimise the risk of geographical miss. In 15% of participants, the posterior ITV-T margin was too tight which could increase the risk of local recurrence (Kim et al., 1995).

Difficulty with the CTV-E was more surprising as it was identified as an issue by only 13% of questionnaire respondents. 41% of RTQA participants contoured the superior border of the CTV-E incorrectly at their first attempt; possible explanations include the change in practice to three different risk-adapted nodal volumes, and variation in the literature on the definition of the superior limit of the para-aortic elective lymph node volume (Choi et al., 2015, Hata et al., 2015, Ouyang et al., 2017). In 22%, the superior border of the CTV-E was too low which may impact the

risk of para-aortic lymph node recurrences, an endpoint in the EMBRACE-II study. 31% of participants did not adequately cover the left lateral para-aortic space, a common site for para-aortic lymph node recurrences (Pano et al., 2015, Takiar et al., 2013); however, other delineation guidelines have not placed the same importance on this issue (Keenan et al., 2018). The majority of our clinicians contour other pelvic tumour types (which often have slightly different elective nodal volumes) in their clinical practice - this may have added to the confusion in selection and delineation of the CTV-E.

Discrepancies between delineation guidelines may also have contributed to difficulty with the CTV-T_LR_{init}. The delineation guidelines most commonly used by questionnaire respondents were the Gyn IMRT consensus guidelines (Lim et al., 2011). As mentioned in Chapter 2 (2.4.1), these define the lateral border of the parametrium as the medial edge of the internal obturator muscle/ischial ramus (i.e. lateral to the pelvic vessels) while the EMBRACE-II protocol defines it as the medial edge of internal iliac and obturator vessels. In our study, discrepancy in the lateral parametrium border was noted for 16% of participants. However, this discrepancy was assessed as having low impact on loco-regional control as this border merges into the CTV-E laterally.

Assessment validity

Like most clinical trials, our RTQA process was limited to PIs although our online programme did allow non-PIs to participate on a voluntary basis. Only 9 non-PIs submitted contours for evaluation; of these, 6 failed at the first attempt. For 5 of the 6 failing non-PIs, the PI at their centre had passed first time suggesting that the contouring performance and learning needs of non-PIs may be different from PIs at any given centre. Ideally, RTQA processes should be extended to include non-PIs but the resources required are considerable.

The use of an online contouring tool which is different from the clinical tools used by participants (i.e. the 'response process' - see Table 3-6) may have contributed to the errors seen. In EMBRACE-I, each centre submitted a "good response" case and a "poor response" case contoured in their own clinical software for review (Kirisits et al., 2015). While this circumvented the issue of unfamiliar software, there may be a different bias in that centres may have chosen their "best" cases. Despite this, 11 of the 28 (39%) centres assessed in the EMBRACE-I quality assurance process had 'major inconsistencies' in external beam contouring and 13 (46%) in IGABT contouring. The impact of unfamiliar contouring software on performance in this study is therefore unlikely to be major.

This accreditation was based on a single accreditation case and the extent to which the same or different errors would occur in other clinical scenarios was not evaluated. The INTERLACE study

of cervical cancer used two accreditation cases (Eminowicz and McCormack, 2015) and reported several differences in the location/magnitude of contouring discrepancies between the cases. In this study we did not assess contouring of the superior border of the CTV-E for “large pelvis” irradiation, another common site for recurrence (Beadle et al., 2010). This border is explicitly defined as the aortic bifurcation in the EMBRACE-II protocol but other contouring guidelines (Small et al., 2008, Group et al., 2010) have specified the L4/L5 interspace (which is usually below the bifurcation) as an acceptable surrogate. Another challenge for RTQA is therefore to improve contouring consistency across a range of cases without increasing the burden on participants or assessors. This is discussed further in Chapter 6.

Automated assessment

Geometric indices such as JCI are frequently used as a measure of contouring adequacy in RTQA programmes and have the advantage of being automatically generated by most contouring software. However, our study showed that of the contours available for JCI analysis, 40-74% passed by expert assessment would have failed by a standardised JCI cut-off depending on the threshold chosen, and up to 37% of failing contours would have been accepted. The SCALOP pancreatic trial RTQA findings were discussed above - the authors found that a high concordance ($JCI \geq 0.7$) between investigator and gold-standard contours predicted a worse outcome (Fokas et al., 2015). The relationship between conformity indices & clinical adequacy is therefore not straightforward. Time-efficient and reliable methods of assessment which can differentiate between clinically important and unimportant discrepancies are required (Vinod et al., 2016a, Jameson et al., 2010).

The reference contour was agreed by consensus, but commonly in RTQA a STAPLE contour (Warfield et al., 2004, Eminowicz et al., 2016b) is produced. The online tool did not have the functionality to calculate a STAPLE contour - a limitation of this study - but other faculty contours were taken into account when assessing participant’s submissions. Hellebust et al. showed a small difference between STAPLE-assessed and expert-contour-assessed conformity indices for brachytherapy (Hellebust et al., 2013), but neither was superior to the other in their prediction of dosimetric impact.

While the documentation of qualitative comments allowed us to identify common and/or clinically relevant errors, the comments were inherently heterogenous as contours were assessed by different clinicians and an objective marking schema was not used (Cox et al., 2019, Setyonugroho et al., 2015). The analysis was retrospective: prospective coding of errors by predicted severity would provide more robust evidence of the significance of specific variations in delineation. A systematic analysis of the dosimetric (or indeed clinical) impact of the

delineation errors identified is beyond the scope of this study – the potential consequences identified in Table 5-2 - Common and/or clinically significant errors seen in first submissions for EMBRACE-II IMRT benchmark case were agreed by the authors' consensus. A checklist, as provided in many clinical skills examinations (Daniels et al., 2014), would be likely to aid the objectivity of the assessment process but the construction of such a checklist is a significant piece of work in itself.

Online CME programme

There was moderate engagement with the optional learning materials in our CME programme. Studies have shown that while online learning is now well-established in continuous professional development for physicians (Cook et al., 2018), completion rates are relatively low especially in large-scale asynchronous courses (Jordan, 2014). Only 28% of our RTQA participants submitted contours on the practice cases. Similar findings were reported by the RAIDs cervical cancer study which conducted online delineation workshops as part of their RTQA process (Rivin del Campo et al., 2017) - only 9/46 (20%) clinicians completed all six contouring exercises. Qualitative investigation is necessary to ascertain the causes of and identify strategies to improve the low engagement - both surveys and in-depth interviews would provide helpful data.

Despite the limited participation in the CME programme, valuable insights were gained. The pre-accreditation questionnaire highlighted discrepancies between clinicians' perceptions of their difficulties and their performance in the accreditation evaluation - 8 of 10 respondents who reported "no particular difficulty" with contouring required revisions to their first submission. Strategies are therefore required to raise individual awareness of actual difficulties.

Quizzes may be a quick way of highlighting key aspects of the protocol and identifying areas of difficulty - the CTV-E quiz question received the lowest score and was also the only question which attracted a comment from a participant. Given this, it makes sense to try to improve their uptake by clinicians – increased publicity, progress monitoring and/or gamification of learning (Looyestyn et al., 2017) may be helpful strategies. The contouring atlas was accessed by most clinicians many times; integrating interactive learning exercises into this resource may enhance its educational value.

Use of learning analytic data has not been reported in radiotherapy educational programmes before. For this programme, learning analytic data has provided useful information to the trial management group about the uptake of different learning materials. The relationship between accessing CME contouring and passing the assessment is likely to be complex. Accessing the CME contouring resource may lead to increased performance in the contouring exercise (i.e. there may

be a learning effect), but it could also be true that clinicians with higher intrinsic motivation for self-improvement (independent of the CME programme) were more likely to access the learning opportunities and to perform well in the test case. Conversely, it is possible that clinicians with a higher level of clinical experience, training and/or confidence may have been (appropriately) less likely to seek out self-directed learning opportunities in the CME programme. Data collected in the pre-accreditation questionnaire about the clinicians' personal experience and confidence were not detailed. Given these limitations, we cannot draw any conclusions about the relationship between accessing CME contouring and delineation performance.

In their systematic review of educational interventions to improve radiotherapy contouring (Cacicedo et al., 2019) discussed in Chapter 2 (Section 2.4.4), Cacicedo et al. note that while onsite, online and blended learning courses have all been shown to be helpful, the impact assessment carried out in the studies was almost exclusively short-term. The authors conclude that *"the most effective teaching methodology/format is unknown and the impact on daily clinical practice is uncertain"* (p.86).

When trying to change practice or introduce new concepts, established mental models (Gentner and Stevens, 2014) of delineation built up by clinicians may have to be challenged repeatedly before meaningful learning takes place (Cook et al., 2010). In the context of assessment in quality assurance, a key challenge is increasing the awareness of participants who are likely to have difficulties, whilst not significantly increasing the burden on the participants who are already competent. Sequential assessment (Pell et al., 2013) and/or increased use of formative assessment (Eva et al., 2016) may be helpful in doing this.

5.5 Conclusion

This study has demonstrated the feasibility and highlighted the challenges of creating an online CME programme to support clinical trial participants (PIs and non-PIs) in RTQA.

Moderate engagement with the optional educational activities in the CME programme was seen, and the pass rate for the accreditation case was relatively low.

Lower performance was seen in ROIs that involved new concepts, a change of practice, and contradictions with other guidelines. The errors identified have formed the basis for the development of further teaching materials. Clinicians seem not be able to fully predict the difficulties they will encounter. Given the difference in performance of PIs and non-PIs (from the

same centre) that was seen, a more detailed exploration of the difference in their performance and learning needs is warranted.

This study has highlighted the need for innovations for teaching and assessing contouring competency including improved methods of assessments and strategies to encourage user engagement and challenge existing mental models. This is the focus of the second half of this thesis and will be discussed further from Chapter 7 onwards.

6 EMBRACE-II brachytherapy quality assurance: contouring assessment

6.1 Introduction

Chapter 1 outlined the technical and conceptual advances in cervix cancer brachytherapy that have led to significantly improved patient outcomes over the last 20 years. MRI guidance and intracavitary/interstitial (IC/IS) brachytherapy have allowed improved visualisation of the target (and avoidance) structures at the time of brachytherapy, and better conformality of the brachytherapy dose around these target structures. This facilitates radiotherapy dose escalation to achieve better local tumour control as well as sparing of organs at risk to reduce toxicity.

When image-guided adaptive brachytherapy was pioneered new target concepts (Table 6-1 & Figure 6-1) were central to its implementation; these were published by Haie-Meder et al. in 2005 (Haie-Meder et al., 2005).

These target concepts have been validated in the multi-institutional retrospective cohort study retro-EMBRACE which reported crude pelvic control rates of 87% at 5 years - an increase of more than 10% over comparable historical cohorts (Sturdza et al., 2016). Further validation of the target concept has since come from the prospective EMBRACE-I trial showing still higher rates of local control and overall survival, especially for stage IIIB/IV tumours where 5-year local control increased to 91-92% compared with 75% in the retro-EMBRACE study (Pötter et al., 2020).

Table 6-1 - Regions of interest and target concepts for cervical cancer image-guided adaptive brachytherapy

Region of interest (ROI)	Abbreviation	Definition
Target structures		
Residual gross tumour volume	GTV _{res}	Residual macroscopic tumour present at the time of brachytherapy Combination of clinical examination findings and high signal area on T ₂ -weighted MRI (or diffusion-weighted MRI)
High-risk clinical target volume	HR-CTV / CTV _{HR}	Macroscopic disease and areas at highest risk of recurrence Includes the GTV _{res} , the remaining cervix tissue and areas of tumour fibrosis ("grey zones"). This is the volume to which the brachytherapy dose is prescribed
Intermediate-risk clinical target volume	IR-CTV / CTV _{IR}	Conceptually: the potential extent of microscopic disease at brachytherapy Practically: the original tumour extent mapped onto the anatomy at the time of brachytherapy after tumour shrinkage has occurred; usually formed from HR-CTV with a 5-15mm margin adapted to anatomic boundaries
Organs at risk		
Bladder		The outer wall of the bladder
Rectum		The outer wall of the rectum
Sigmoid		Contoured superiorly to the rectum until at least 2cm above the target structures
Bowel		Loops of small or large bowel at the level of, or less than 2cm superior to, the target structures

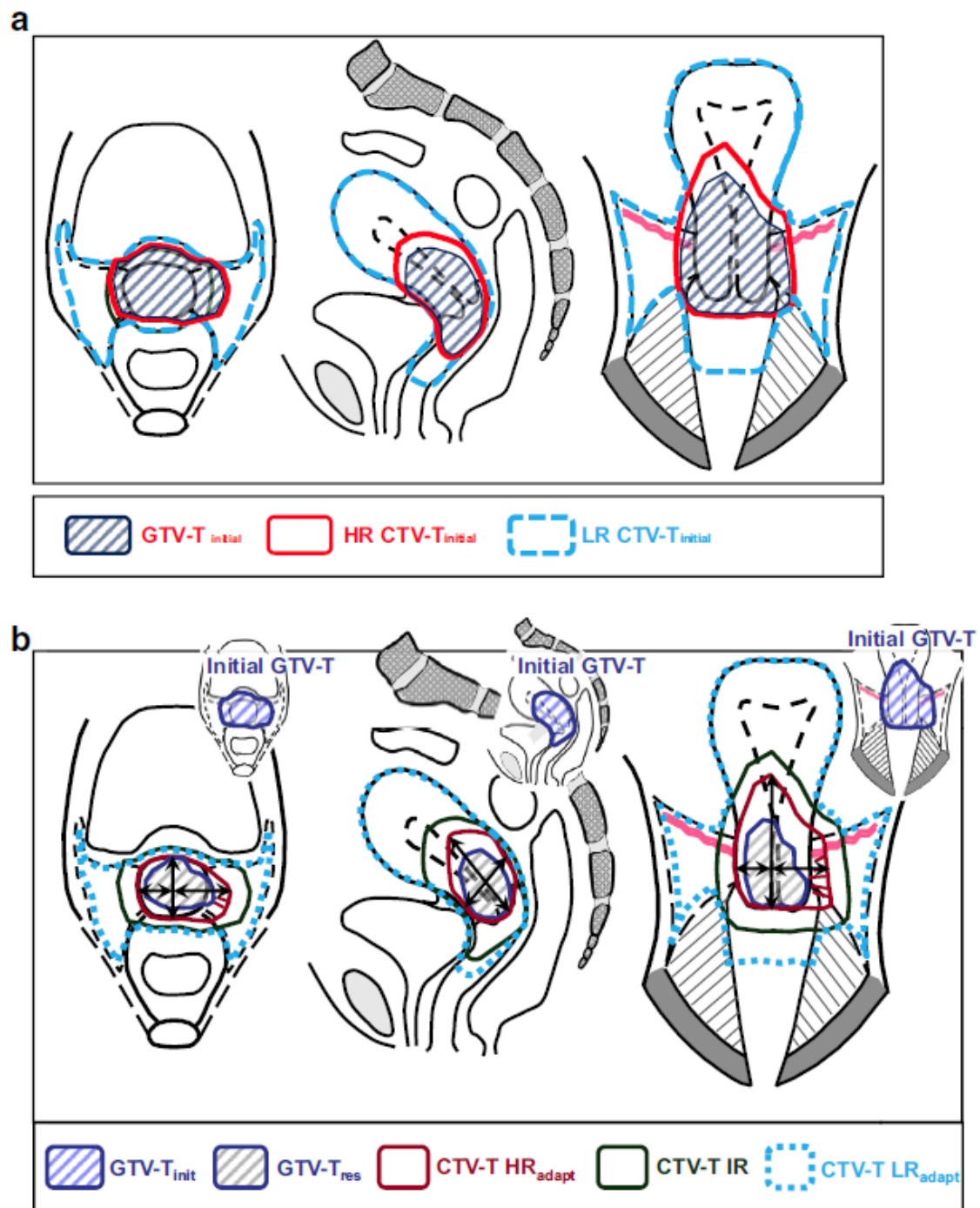


Figure 6-1 - Target concepts in image-guided adaptive brachytherapy for cervix cancer. Reproduced from EMBRACE-II protocol (Pötter et al., 2016a)

(a) illustrates the tumour at the time of diagnosis (grey shading) with areas of possible microscopic tumour spread (see Chapter 5).

(b) illustrates the tumour at the time of brachytherapy (residual GTV in blue/grey shading). The regions most at risk of microscopic spread are: (i) the high-risk CTV (red); a combination of the residual GTV, any normal cervix tissue and fibrotic tissue ("grey zones") seen in areas where tumour has receded (ii) the intermediate-risk CTV (green) which represents the extent of the tumour prior to treatment superimposed on the anatomy at the time of brachytherapy.

Previous studies have shown significant inter-clinician variation in cervix cancer brachytherapy contouring, especially for the residual gross tumour volume (GTV_{res}) and the intermediate-risk clinical target volume (IR-CTV / CTV_{IR}) (Petric et al., 2013, Petric et al., 2008). Contouring of the high-risk clinical target volume (HR-CTV / CTV_{HR}), which is the volume to which the dose is prescribed^{ix}, is generally more consistent (Petric et al., 2013) but inter-clinician variation is still seen. This variation in contouring contributes to significant dosimetric uncertainty - between 6% (Bell et al., 2020) and 35% (Vinod et al., 2017) of the prescribed dose at brachytherapy. Contouring is considered to be the second largest source of uncertainty in calculating the dose to the target for brachytherapy, second to intra- and inter-fraction organ motion/filling uncertainties (Tanderup et al., 2013).

The steep dose gradient of brachytherapy (see Chapter 1, Figure 1-5) means that contouring errors are more likely than for external beam radiotherapy to translate into an underdose to the tumour (resulting in an increased risk of local recurrence), or an overdose to organs at risk (resulting in an increased risk of severe toxicity):

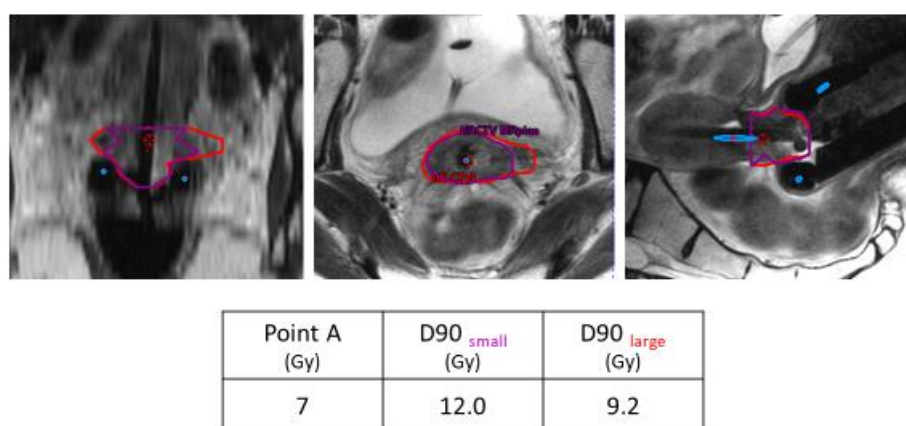


Figure 6-2 - Impact of contouring on the reported dose to 90% of the HR-CTV ("D90"). Under-contouring can lead to falsely high dose. Figure courtesy of Dr Li-Tee Tan.

As explained in Chapter 2, high-quality contouring is an important component of good clinical outcomes. Uniquely to radiotherapy clinical trials such as EMBRACE-II, consistent contouring is also vital for the validity of analysis of dose and volume effect relationships. The dose-volume effects seen in the retro-EMBRACE and EMBRACE-I studies are not just of academic interest - they have been important in defining radiotherapy treatment planning aims and objectives for routine clinical practice (Tan et al., 2019b, Tanderup et al., 2020). Contouring variation in EMBRACE-II could loosen or obscure associations between dosimetric data and clinical outcomes and thereby

^{ix} Brachytherapy dose is prescribed to 90% of the HR-CTV volume or the "HR-CTV D90"

have implications for brachytherapy planning parameters in the clinic. Quality assurance to ensure consistent contouring within the EMBRACE-II trial is therefore doubly important.

In contrast to the IMRT component of EMBRACE-II, no new target concepts were implemented for brachytherapy. Therefore those centres who had completed the EMBRACE-I quality assurance were excused from brachytherapy contouring quality assurance and accredited to EMBRACE-II once they had completed the IMRT RTQA. This chapter reports the results of the EMBRACE-II brachytherapy contouring assessment for those centres who were new to the EMBRACE trial group.

The aims of this chapter are to:

- Examine the impact of case variation within contouring quality assurance
- Compare and contrast performance and errors with external beam radiotherapy
- Compare the contouring performance of these entrants to the EMBRACE-II trial with previous cohorts for the same target volumes
- Identify potential causes of contouring errors

My contribution

This study is the result of collaborative research within the EMBRACE-II trial group. Collaborators and their affiliations are listed below in Appendix Table A.6-1. The EMBRACE-II quality assurance process was designed by the trial management group. I was involved in assessing the contouring (see below) and have collated, analysed and written up the data below. I am grateful for the assistance of Dr Hatem Helal in coding an initial MATLAB import of the contour data from the Addenbrooke's Contouring Tool.

6.2 Methods and materials

6.2.1 Participants and timeline

Participating centres were required to have experience of MRI-guided IGABT in clinical practice prior to accreditation, treat more than 10 patients per year in their centre, and routinely be using combined intra-cavitary & interstitial brachytherapy.

EMBRACE-II brachytherapy quality assurance comprised results from the centre compliance questionnaire, contouring assessment, and test-patient review based on dosimetric data (as well as contouring screenshots) from five clinical cases treated as per EMBRACE-II protocol.

Contouring assessment was conducted using benchmark cases on web-based simulator software as described in the previous chapter (Addenbrooke's Contouring Tool - see Figure 5-1). The 'test patient' review is ongoing.

66 centres passed the EBRT contouring and planning quality assurance. 17/66 centres had previously participated in the EMBRACE-I trial (and quality assurance) and so were excused from the brachytherapy contouring and test patient processes. For the remaining 49 centres new to the EMBRACE trial group, one clinician from each centre - the principal investigator - was required to pass the brachytherapy contouring quality assurance before the centre progressed to the test patient phase.

Assessment of the initial tranche of 5 submissions took place in July 2017, and thereafter were conducted in batches; the final deadline for new submissions or re-submissions was 31st August 2018 and assessments were completed in November 2018.

6.2.2 Accreditation cases

There were two accreditation cases. Clinicians contoured the GTV_{res}, HR-CTV, IR-CTV, rectum, sigmoid, bladder and bowel on MRI images with the brachytherapy applicators in situ. The Addenbrooke's Contouring Tool contained a case history including diagrams of clinical examination findings at diagnosis and at brachytherapy, and pre- and post-EBRT multiplanar diagnostic MRI imaging.

Contouring guidance for brachytherapy within the protocol was concise (3 pages with 3 figures) and referenced the more extensive guidance in the ICRU 89 report (ICRU, 2013) and GEC-ESTRO recommendations (Haie-Meder et al., 2005). These resources contain 2D images with expert consensus target and organ at risk volumes for multiple cases. Clinicians treating cervix cancer with brachytherapy should be familiar with these documents although their understanding was not checked prior to accreditation.

Case 1

Case 1 was a patient with stage T3bN0M0 squamous cell carcinoma at diagnosis with an exophytic tumour involving the distal left parametrium (i.e. extending to the pelvic sidewall) and proximal right parametrium. All fornices of the vagina were involved, with extension of the tumour 40mm along the anterior vagina:

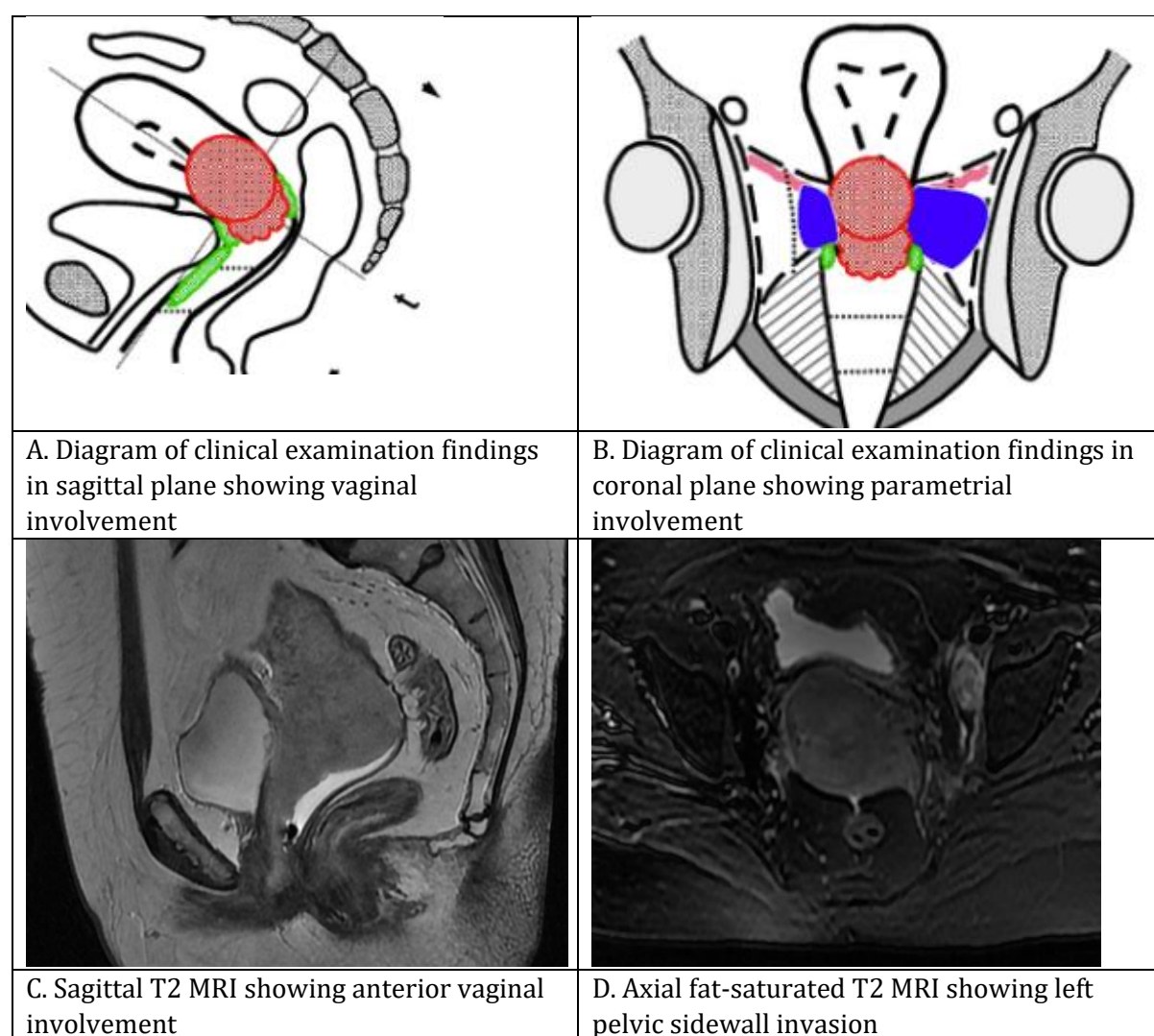


Figure 6-3 - Case 1 Clinical and MRI findings at diagnosis

After EBRT and chemotherapy there was partial regression of the exophytic tumour with residual endophytic disease at the cervix. The disease extent at brachytherapy is shown in Figure 6-4. A Vienna-I ring applicator and 4 interstitial parametrial needles were inserted for brachytherapy treatment.

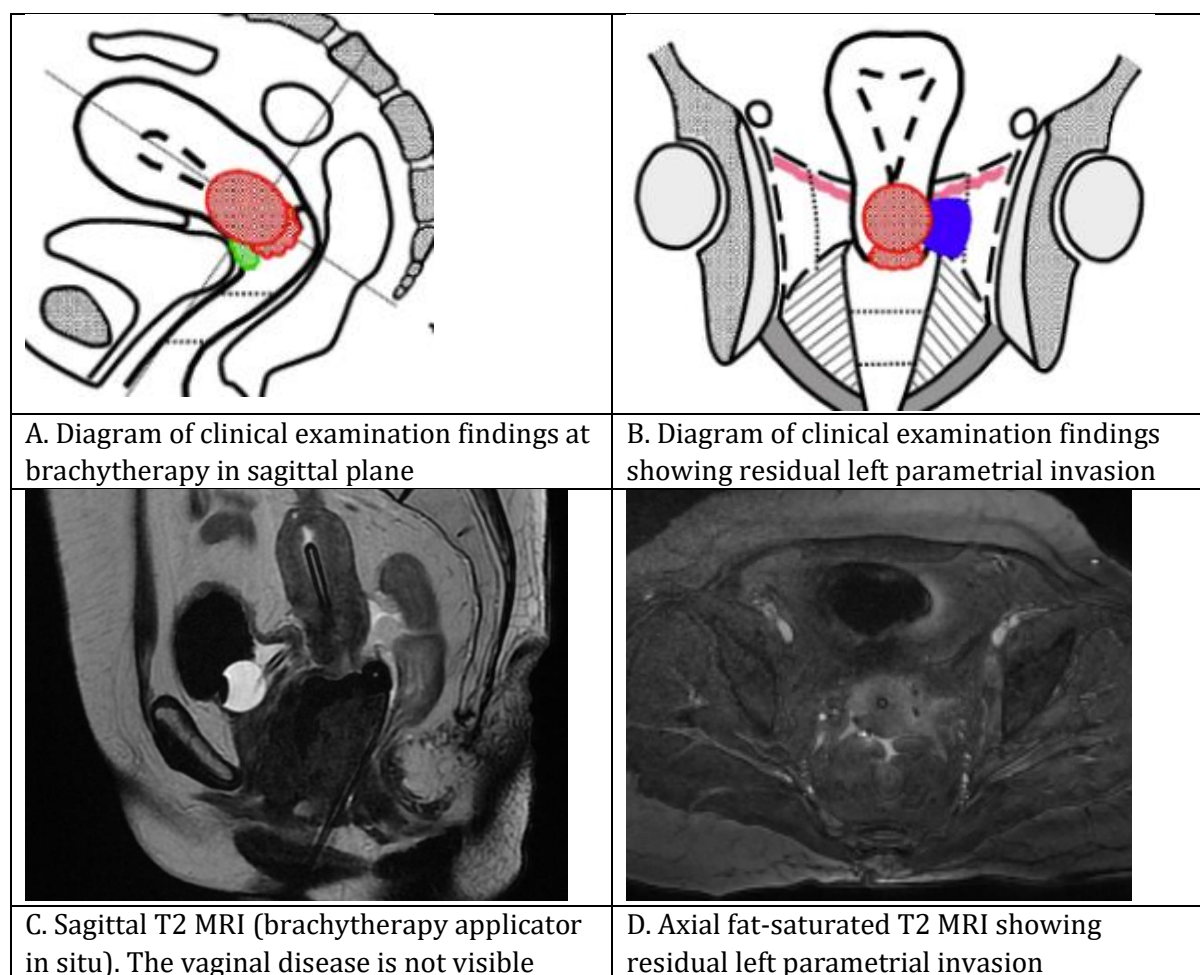


Figure 6-4 - Case 1 clinical and MRI findings at brachytherapy

Case 2

Case 2 was a stage T3bN0M0 patient with distal right parametrial (up to the pelvic sidewall) and proximal left involvement at diagnosis; there was no vaginal involvement at diagnosis:

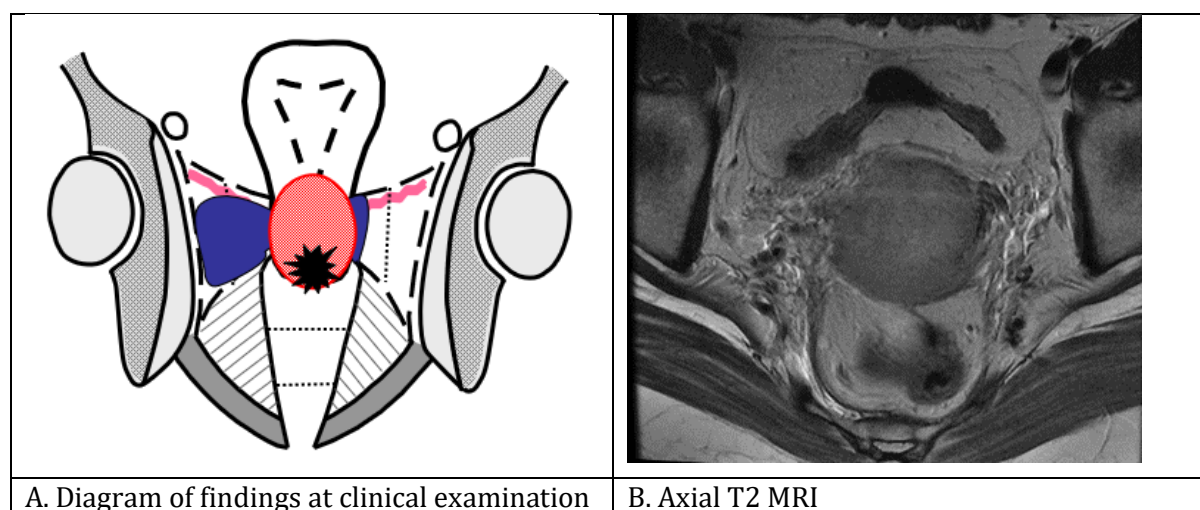


Figure 6-5 - Case 2 clinical and MRI findings at diagnosis

After EBRT residual disease was seen on the anterior lip of the cervix, but no residual parametrial invasion was palpated on clinical examination or seen on MRI. An intra-cavitary Vienna-I ring applicator was inserted for brachytherapy treatment, but no interstitial needles were inserted.

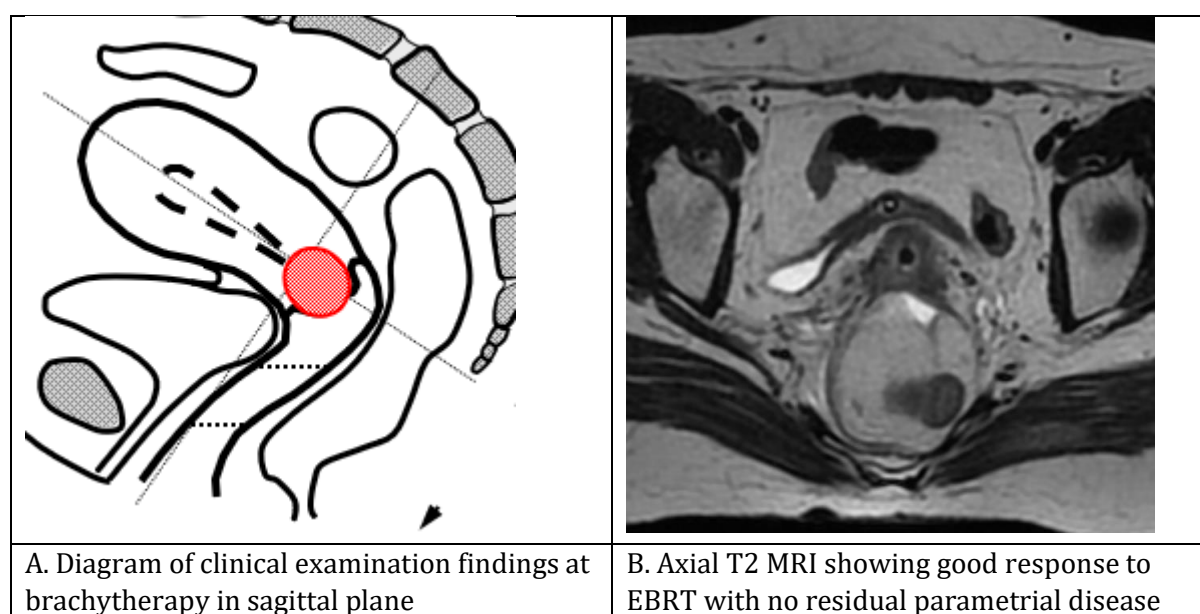


Figure 6-6 - Case 2 clinical and MRI findings at brachytherapy

6.2.3 Assessment

Scoring

The reference contour was created by consensus at a meeting of the Trial Management Group in July 2017. The first 5 submissions were assessed jointly by 3 oncologists and 2 medical physicists

(RP, NN, MS, AS, KT) at a video-conferenced meeting in July 2017. Each region of interest (ROI) was compared to the consensus reference and was scored between 1 and 10 (Table 6-2) - this differed from EBRT as the scale was adapted from the brachytherapy quality assurance in EMBRACE-I (Kirisits et al., 2015) to maintain continuity. Assessors were not blinded to the centre that they were assessing. At this meeting the principles for assigning scores were decided – Failing scores (≤ 5) were assigned to contours with errors demonstrating a flaw in conceptual understanding that was likely to impact clinical outcome and/or have a dosimetric impact i.e. a major deviation. Scores of 6 & 7 were assigned to contours with minor deviation and scores of 8 or above were given to good or excellent contours with no significant deviations. Figure 6-9, Figure 6-10 and Figure 6-11 in the results section show examples of how the scoring system was applied to various contouring errors.

Table 6-2 - Assessment scale for EMBRACE-II brachytherapy regions of interest; adapted from EMBRACE-I brachytherapy quality assurance

Score	Rating
10	Excellent, no significant disagreement.
8-9	Good, only minor disagreement referable to inter-observer variability.
6-7	No major violation of the study protocol but variation from reference contour noted. Comment given.
4-5	Poor, clear violations of the study protocol. Explanation given.
2-3	Very poor. Explanation given.
1	Very poor, total disagreement. Explanation given.

Subsequent submissions were assessed remotely by at least two oncologists (two of AS, MS, RP, SLD) using the previously agreed principles. Disagreements or queries were discussed with the wider group to finalise a consensus score. The time taken for clinicians to contour or assessors to mark the contours or was not recorded.

Accreditation and feedback

Clinicians required a passing score (of ≥ 6) on every ROI on both cases in order to pass this stage of the accreditation. Clinicians were given qualitative feedback for every ROI where they scored ≤ 7 and the most important points were emphasised in an overall comment.

Clinicians who failed the initial submission were encouraged to re-submit after reviewing the relevant parts of the EMBRACE-II protocol and ICRU 89 document.

6.2.4 Data analysis

As for the IMRT contouring quality assurance in Chapter 5, a retrospective analysis of first-attempt submissions was conducted. The Jaccard conformity index (JCI) was calculated using MATLAB (The Mathworks Inc., 2018) for each participant ROI against the reference contour and was compared to the expert-assigned score. Missing contours or contours that had not been scored by the assessors were excluded from the conformity index analysis. Qualitative feedback comments for ROIs with scores less than or equal to 7/10 (i.e. 'fair' or worse) were analysed and collated in the same way as for IMRT (Chapter 3). Analysis of the severity of clinical impacts was conducted by a single expert (LTT).

A pairwise comparison of clinicians' performance for each ROI in the two cases was performed to explore whether participants' errors in ROIs carried across both cases.

6.2.5 Ethical approval

The EMBRACE-II trial has ethical approval and is sponsored by the Medical University of Vienna. Consent for analysis of participant data for the purposes of education and research was obtained at the point of entry to the delineation tool. All centres actively recruiting patients also have national and local ethical approval.

6.3 Results

6.3.1 Overall performance

Forty-nine clinicians submitted contours on two cases, giving a total of 98 submissions. 4/49 (8%) passed at the first attempt and a further 27 (62%) on resubmission after individualised feedback. 11 clinicians (22%) passed at a subsequent attempt and 7 (14%) did not re-submit after one or more failures, and therefore did not progress further with the trial accreditation. Overall, 189 separate contour assessments were performed.

The majority of submissions (59/98 = 60%) failed on more than one ROI, as seen in Figure 6-7. 30/98 (31%) failed on three or more out of the seven assessed.

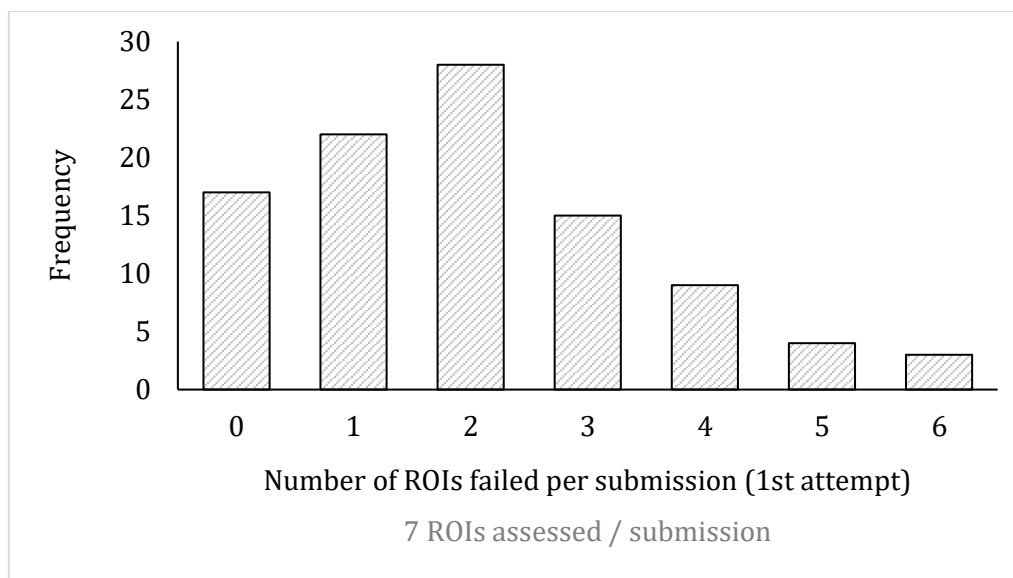


Figure 6-7 - Number of regions of interest (ROIs) failed per 1st attempt submission. Forty-nine clinicians submitted 2 cases for a total of 98 submissions.

Average scores per ROI (for the first attempt) are shown in Figure 6-8. For target volumes, clinicians scored highest on the HR-CTV with an average score of 6.9/10. Average scores were lower for the GTV_{res} (5.5/10) and IR-CTV (5.9/10):

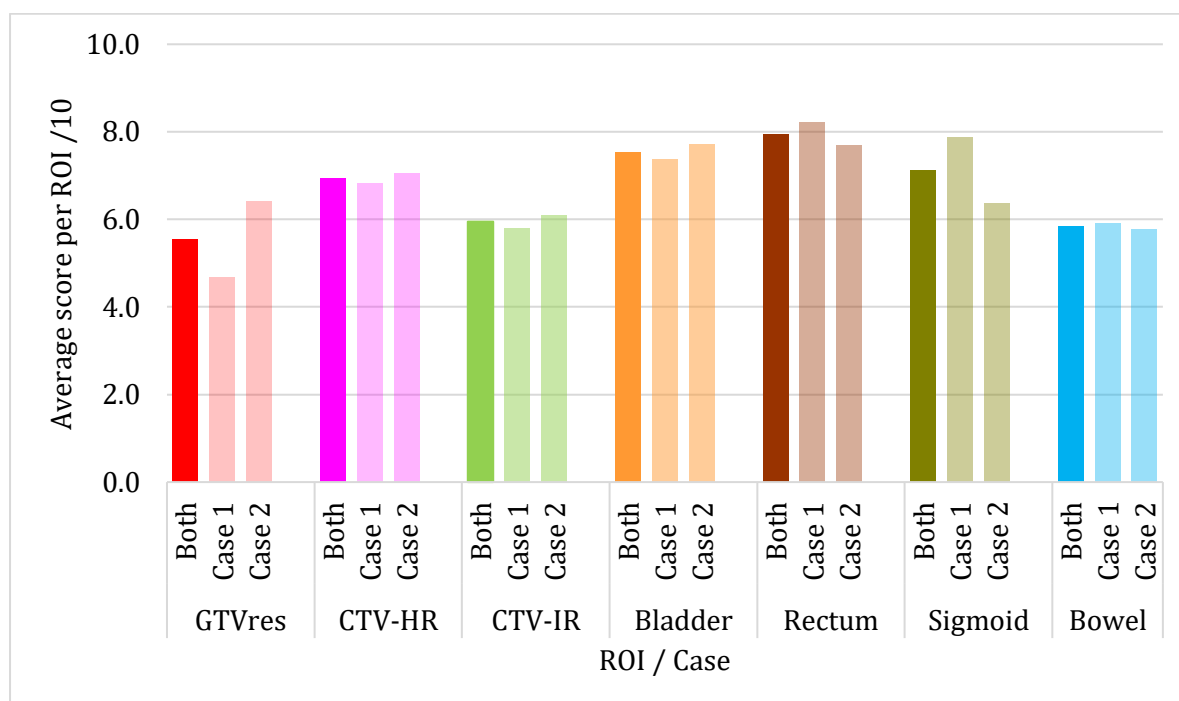


Figure 6-8 - Average scores (1st attempt) per region of interest for the EMBRACE-II Brachytherapy contouring quality assurance

6.3.2 Accreditation errors

Common contouring errors were seen in most ROIs - Table 6-4 lists the common and/or clinically significant errors. A total of 47 distinct contouring errors were coded.

Performance varied by case in some ROIs (Figure 6-8). Scores for the GTV_{res} were notably lower on Case 1 (average = 4.7) than on Case 2 (average = 6.4). The majority (57%) of participants over-contoured the GTV_{res} in Case 1 - this was reflected in the larger median participant volume compared to the reference (see Appendix Table A.6-3: 31cm³ versus 13cm³). 12 clinicians (24%) failed the GTV_{res} across both cases (Table 6-3) - the highest proportion for any ROI.

Table 6-3 Pairwise comparison of failure rates per clinician and per case, listed by region of interest (ROI). A pass on each ROI was defined as a score of ≥ 6 .

ROI	Pass both	Fail both	Fail only Case 1	Fail only Case 2
GTV _{res}	17 (35%)	12 (24%)	18 (37%)	2 (4%)
CTV-HR	32 (65%)	6 (12%)	6 (12%)	5 (10%)
CTV-IR	19 (39%)	11 (22%)	12 (24%)	7 (14%)
Bladder	40 (82%)	3 (6%)	4 (8%)	2 (4%)
Rectum	46 (94%)	1 (2%)	1 (2%)	1 (2%)
Sigmoid	25 (51%)	5 (10%)	1 (2%)	18 (37%)
Bowel	18 (37%)	12 (24%)	8 (16%)	11 (22%)

For the HR-CTV there was less of a pattern of errors across cases. Under-contouring in the axial plane was more common in Case 2 (14% vs 31%).

Clinicians failed the IR-CTV on both cases nearly as frequently (22%) as the GTV_{res} (24%). Despite similar average IR-CTV scores on both cases the errors differed. Several errors on Case 1 related to the disease initially extending down the anterior vagina - an area which should be covered by the IR-CTV for Case 1 but was not relevant in Case 2.

Errors also varied by case for the organs at risk. For the bladder, 5 (10%) clinicians delineated urine rather than bladder wall (in some areas) in both cases - for 3 of them this was severe enough

for them to fail. A further 3 clinicians failed by making this error solely for Case 1 and 2 clinicians for Case 2.

Errors for the sigmoid and rectum on case 2 were linked as due to the unusual right-sided sigmoid clinicians often confused the two. This was reflected in the median participant volumes (Appendix Table A.6-3).

Table 6-4 - Common and/or clinically important delineation errors seen in the EMBRACE-II brachytherapy target volume contouring quality assurance, per ROI

ROI	Error	Under/ Over contouring	Frequency (overall)	%	Freq. Case 1	%	Freq. Case 2	%	Clinical impact
GTV_{res}	GTV not focussed on T2 MRI high signal; includes normal cervix	Over	42	43%	29	59%	13	27%	Low - included in high dose region
	Areas of residual tumour seen on examination missed	Under	25	26%	16	33%	9	18%	Low - usually small & included in high dose region
CTV-HR	Axially too tight	Under	22	22%	7	14%	15	31%	High - area of high dose gradient
	Upper cervix flat (should be dome-shaped)	-	20	20%	12	24%	8	16%	Low - included in high dose region regardless
	Missed tumour in vagina	Under	14	14%	12	24%	2	4%	
	Superior border too high	Over	13	13%	7	14%	6	12%	
	Superior border too low	Over	10	10%	3	6%	7	14%	
CTV-IR	Vaginal limit too high (too short)	Under	17	17%	16	33%	1	2%	Low - IR-CTV dose is usually adequate if HR-CTV dose is adequate
	Should extend 2cm below cervix into all vaginal fornices	Under	43	44%	16	33%	27	55%	
	Packing contoured in addition to vaginal wall	Over	17	17%	17	35%	0	0%	
	Volume extended outside uterine tissue even though uninvolved at diagnosis	Over	30	31%	17	35%	13	27%	
	Lateral parametrium too tight	Under	26	27%	8	16%	18	37%	
	Included bladder wall / ureter	Over	11	11%	8	16%	3	6%	
	Ring applicator not excluded	Over	29	30%	12	24%	17	35%	

Table 6-5 - Common and/or clinically important delineation errors seen in the EMBRACE-II brachytherapy organ at risk contouring quality assurance, per ROI

ROI	Error	Under/ Over contouring	Frequency (overall)	%	Freq. Case 1	%	Freq. Case 2	%	Clinical impact
Bladder	Bladder wall missed i.e. urine delineated	Under	23	23%	18	37%	5	10%	Moderate - underestimates dose to bladder wall (may increase toxicity)
	Urethra included in bladder volume - should be contoured as a separate structure	Over	15	15%	6	12%	9	18%	Low - significance of sub-volume doses to be established
Rectum	Inferior border too low / anal canal included	Over	24	24%	8	16%	16	33%	Low - likely to be in low dose region
	Superior border too low	Under	10	10%	2	4%	8	16%	Moderate - can lead to underestimation of dose
Sigmoid	Sigmoid mis-labelled as bowel	Under	25	26%	2	4%	23	47%	Moderate - can lead to underestimation of dose if intra-fraction variation in contouring
	Missed slices superiorly	Under	15	15%	10	20%	5	10%	Low - likely to be in low dose region
Bowel	Bowel bag (not loops) delineated	Over	20	20%	16	33%	4	8%	Low - volume of interest is the 2cc receiving highest dose (nearest applicator)
	Non-bowel structure included	Over	31	32%	8	16%	23	47%	Depends on structure and proximity to applicator
	Some slices not contoured (e.g. just first few nearest uterus contoured)	Under	20	20%	13	27%	7	14%	Low - volume of interest is the 2cc receiving highest dose (nearest applicator)

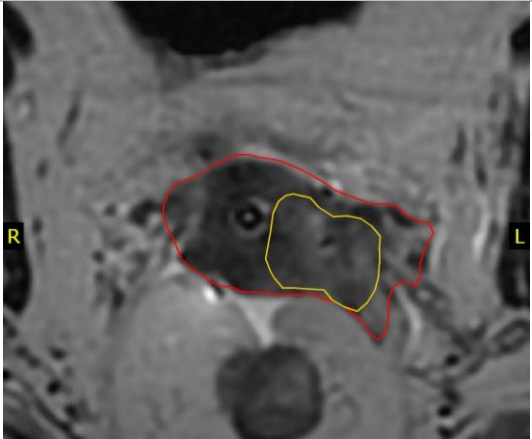
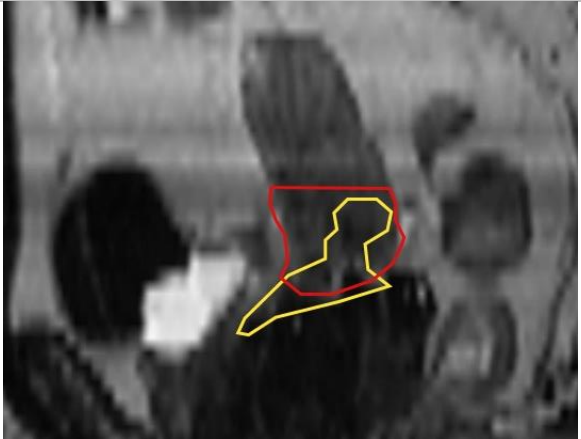
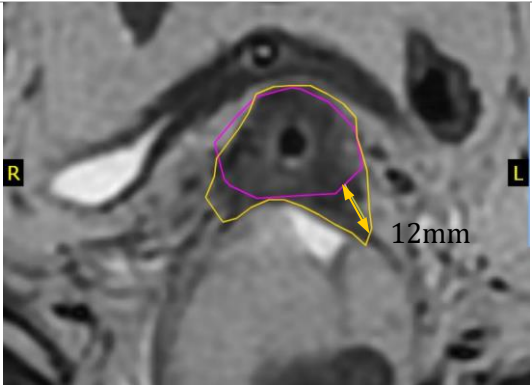

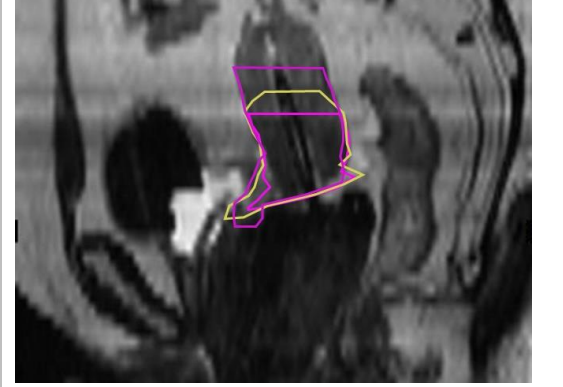
	
<p>A. Case 1 - GTV_{res} not focussed on T2 MRI high signal; includes normal cervix. Scored 2/10.</p>	<p>B. Case 1 - Areas of residual tumour seen on examination missed in GTV_{res}. Scored 5/10.</p>
	
<p>C. Case 2 - HR-CTV axially too tight. Scored 3/10.</p>	<p>D. Case 1 - Missed tumour in vagina. Scored 5/10. (The upper cervix is flat but no penalty for this).</p>
	
<p>E. Case 1 - HR-CTV superior border 2 slices too low & too high. Both scored 6/10.</p>	

Figure 6-9- Example contouring errors and associated scores for the GTV_{res} and HR-CTV. Yellow = gold standard contour; red = participant GTV_{res}; magenta = participant HR-CTV.

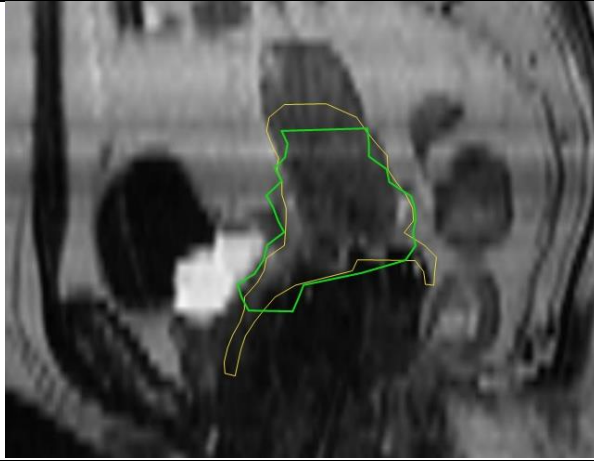
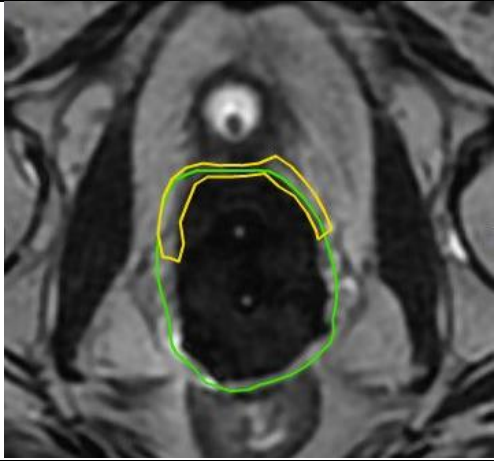
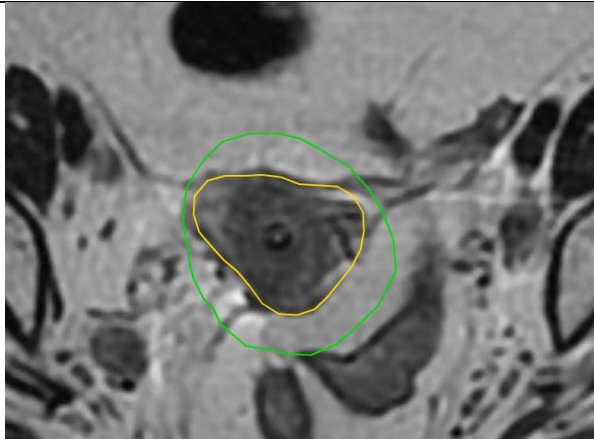

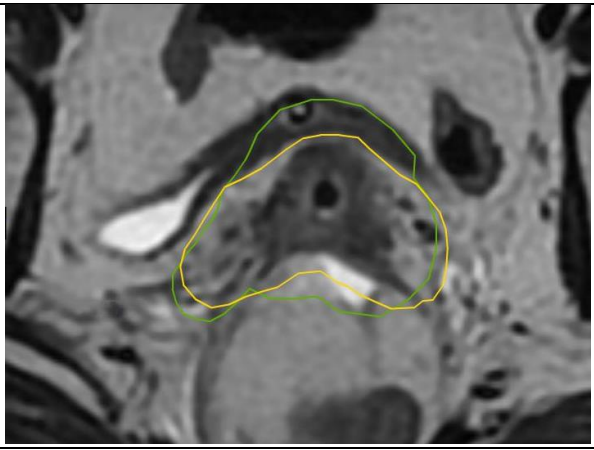
	
<p>A. Case 1 - CTV-IR vaginal limit does not include initial disease extent. Scored 5/10. Posterior fornix not included but no further penalty.</p>	<p>B. Case 1 - Packing contoured in addition to vaginal wall. Scored 7/10.</p>
	
<p>C. Case 2 - Extra-uterine tissues included uninvolved at diagnosis. Scored 6/10.</p>	<p>D. Case 2 - Right lateral parametrium too tight: does not include initial disease extent. Scored 4/10.</p>
	
<p>E. Case 2 - Included bladder wall though uninvolved at diagnosis. Scored 5/10.</p>	

Figure 6-10 - Example contouring errors and associated scores for the IR-CTV. Yellow = gold standard contour; green = participant IR-CTV.

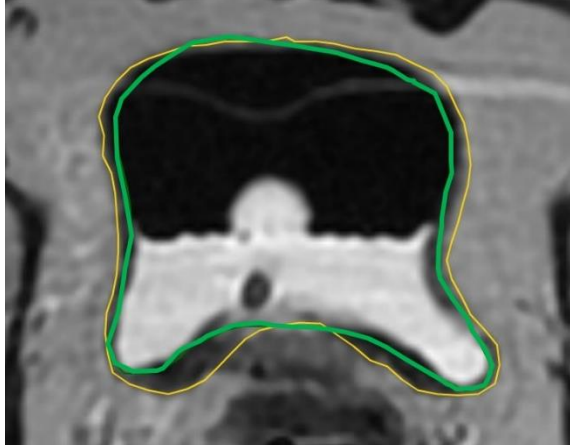
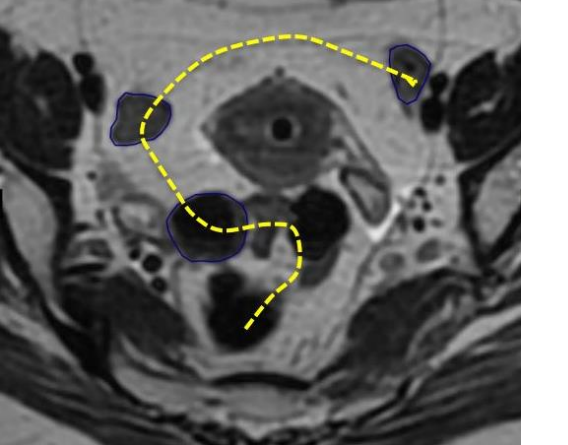
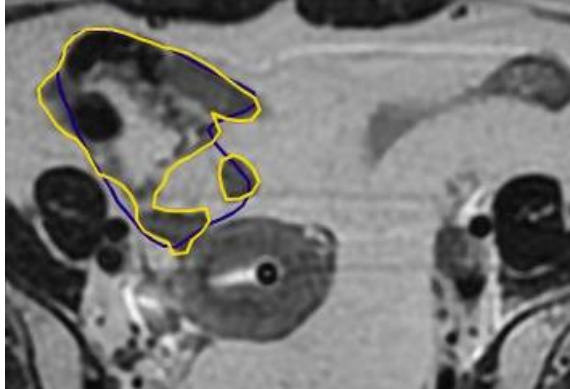
	
<p>A. Case 1 - Bladder (in green) outer wall missed i.e. urine/inner wall contoured. Scored 3/10.</p>	<p>B. Case 2 - Sigmoid mis-labelled as bowel (in blue). Consensus sigmoid path yellow dotted line. Scored 4/10.</p>
	
<p>C. Case 1 - Bowel bag (in blue) instead of individual loops contoured. Scored 5/10.</p>	

Figure 6-11 - Example contouring errors and associated scores for selected organs at risk (see also Table 6-4). Yellow = gold standard contour; green = participant bladder; blue = participant sigmoid(B)/small bowel(C).

6.3.3 Conformity index analysis

Data were available for the conformity index analysis for 87/98 whole submissions (89%) due to second and subsequent submissions overwriting initial submissions in an early version of the Addenbrooke's Contouring Tool. 7 individuals did not contour bowel and 2 individuals omitted sigmoid for case 2.

The JCI per ROI is displayed in Figure 6-12 and

Figure 6-13, as well as Appendix Table A.6-2. Descriptive statistics for volumetric data are displayed in Appendix Table A.6-3.

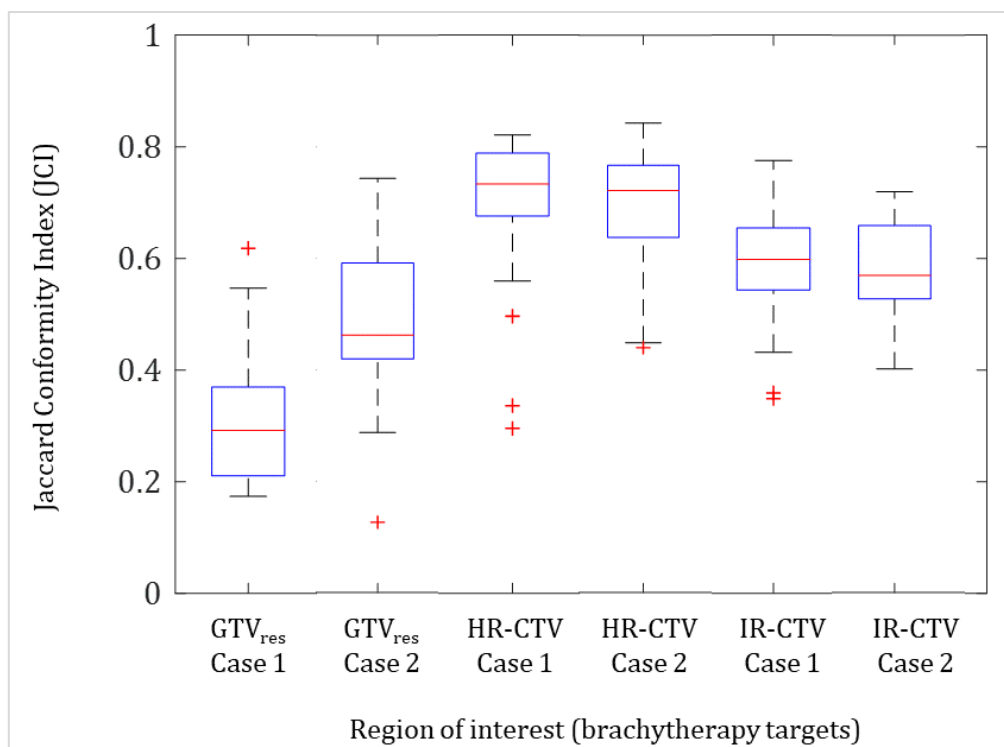


Figure 6-12 Boxplot of Jaccard Conformity Index per target ROI in cases 1 & 2

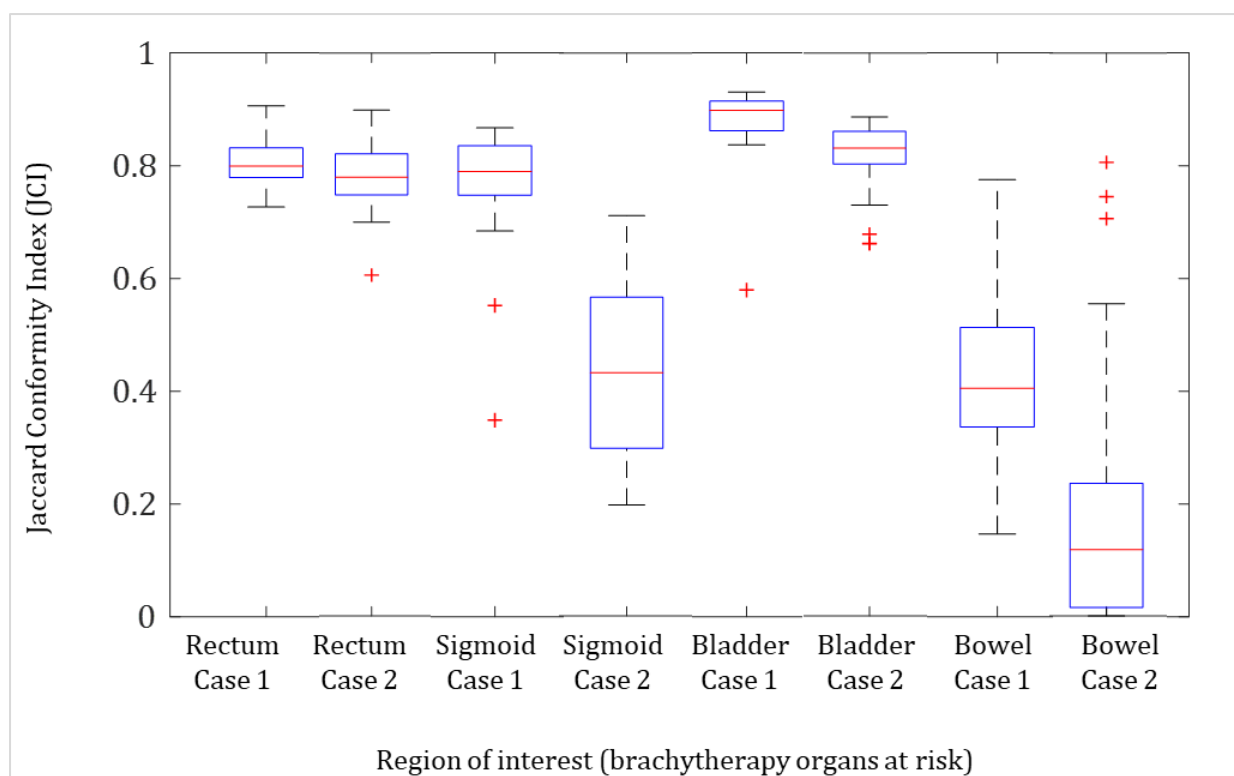


Figure 6-13 Boxplot of Jaccard Conformity Index per organ at risk ROI in cases 1 & 2

Table 6-6 indicates the pass rates for contours that passed or failed manual assessment if an automated JCI cut-off were to be applied across both cases. For the HR-CTV and Sigmoid,

automated conformity index pass/fail thresholds resulted in large distinctions between manually passing and failing groups. For example, an HR-CTV cut-off of 0.7 would detect 52/68 (76%) passes (“true-positives”) and only include 2/19 (10%) fails (“false-positives”). For the other ROIs automated cut-offs either included a lower proportion of true passes or a higher proportion of true fails.

Table 6-6 - Automatic pass classification rates using Jaccard Conformity Index (JCI) vs manual assessment for selected cutoffs - per ROI

ROIs	JCI cut-off	Expert-assessed as “Pass” (true-positive)		Expert-assessed as “Fail” (false-positive)	
Overall		430		152	
	0.6	288	67%	63	40%
	0.7	237	55%	37	23%
	0.75	180	42%	29	18%
GTV _{res}	n = 87	50		37	
	0.6	10	20%	2	5%
	0.7	2	4%	0	-
	0.75	0	-	0	-
HR-CTV	n = 87	68		19	
	0.6	65	96%	12	63%
	0.7	52	76%	2	10%
	0.75	28	41%	1	5%
IR-CTV	n = 85	51		34	
	0.6	28	55%	10	30%
	0.7	10	20%	0	-
	0.75	2	4%	0	-
Rectum	n = 85	80		5	
	0.6	80	100%	5	100%
	0.7	79	99%	5	100%
	0.8	34	43%	2	40%
	0.9	1	1%	0	-
Sigmoid	n = 85	60		25	
	0.6	43	72%	2	8%
	0.7	37	62%	2	8%
	0.8	17	28%	1	4%
	0.9	0	-	0	-
Bladder	n = 86	74		12	
	0.6	73	99%	12	100%
	0.7	71	96%	11	92%
	0.8	68	92%	8	67%
	0.9	18	24%	1	8%
Bowel	n = 80	48		32	
	0.6	8	16%	2	6%
	0.7	4	8%	1	3%
	0.8	1	2%	0	-
	0.9	0	-	0	-

Analysing the cut-offs for the GTV_{res} and IR-CTV by individual case did not produce a major change in the proportions of true and false positives.

6.4 Discussion

These results show that practicing clinicians frequently make errors in applying IGABT target and organ at risk contouring concepts to simulated cases, more than 10 years after these concepts were introduced into clinical practice.

Pass rate and assessment validity

This summative assessment of contouring competency was conjunctive (Ben-David, 2000) (see Section 3.7.1) - i.e. clinicians were required to pass every ROI on both cases to pass the assessment overall. Lower pass rates are seen with conjunctive standards as opposed to compensatory standards (Ben-David, 2000), where good performance in one domain/scenario can compensate for poor performance in another. A conjunctive standard seems appropriate in radiotherapy contouring where contouring the bladder wall correctly should not compensate for missing the tumour. Nevertheless the pass rate of 4/49 (8%) of clinicians was surprising given their clinical experience and that no new concepts were introduced in the trial.

Conjunctive assessment means that measurement error (e.g. assessor rating stringency) in one domain (i.e. one ROI, for example the residual GTV or bowel) could have a significant impact on the overall pass rate, which cannot be ruled out in this study. From the trial group's perspective, an overly stringent assessment may be preferable to a lax one where errors are uncorrected and persist within the trial. Assessment validity with reference to radiotherapy quality assurance was discussed in Chapter 3. Modifications that could have increased the validity of this assessment include: blinding, assessor training, fewer assessment categories with more detailed anchor statements (for global assessment), checklists (for itemised assessment as discussed in Chapter 5), compensatory marking within ROIs across multiple cases, re-contouring on a separate case after failing, and comparison against multiple rather than single reference contours (to illustrate to assessors where variation is acceptable). Unfortunately, most of these adaptations require increased resources and so even if radiotherapy trial groups were aware of these steps they may not be feasible.

The ICRU and EMBRACE-II guidance could be supplemented with a 3-D contouring atlas (Gillespie et al., 2017) and more detailed instructions (coaching against these errors) and this should form part of future work. Indeed adapting protocols or contouring guidance to the errors seen should form part of the routine 'life-cycle' of such guidance (Eminowicz et al., 2016a).

Variation by regions of interest and case

In keeping with the previous literature, the HR-CTV was contoured more consistently with less inter-observer variation. In an early study, Dimopoulos et al. studied the agreement in GTV_{res}, HR-CTV and IR-CTV between two radiation oncologists from Vienna and Paris (Dimopoulos et al., 2009). They found a statistically significant difference of approximately 20cm³ between the two observers' IR-CTV volumes but no difference for the GTV_{res} and HR-CTV volumes. Conformity indices (denoting geometric overlap) ranged between 0.5-0.7 and were deemed satisfactory by the authors when compared to the lack of agreement between inexperienced practitioners (Kelly et al., 2006). This study was unlikely to be representative of applicability of the concept by a wider audience of clinicians, as the two radiation oncologists were from institutes that had jointly crafted the IGABT concepts and written the GEC-ESTRO guidelines, and so are likely to have an unusually high level of shared conceptual understanding.

Petric et al. examined the target contouring variation between 10 radiation oncologists against reference consensus contours (expert consensus and STAPLE algorithm generated) (Petric et al., 2013) over 6 cases. All clinicians participating in this study had significant experience in IGABT, many from the EMBRACE group, so this was still a selected group. There was better geometric agreement between the participants' contours and reference contours for the HR-CTV (mean volumetric conformity indices (equivalent to the JCI) 0.72-0.76, mean delineation distance $3.8 \pm 3.4\text{mm}$) than for the GTV (0.58-0.59, $4.2 \pm 3.5\text{mm}$) and IR-CTV (0.68-0.77, $5.2 \pm 5.6\text{mm}$). The authors concluded that the *"HR-CTV may be considered most robust volume for dose prescription and optimization in cervix cancer IGABT"*. These findings were echoed in the EMBRACE-I trial dummy run (Kirisits et al., 2015), where out of the 28 centres participating, fewer centres' HR-CTV delineation required correction (8/28, 5 major) than the GTV_{res} (10/28, all major) or IR-CTV (14/28, proportion major not specified).

The mean conformity for this cohort was similar to Petric et al. for HR-CTV (0.73 vs 0.72-76), but lower for GTV_{res} (0.40 vs 0.58-59) and IR-CTV (0.59 vs 0.68-0.77). Comparison between these figures may be confounded by case selection and reference contouring, but there may have been a meaningful difference in the ability of the two cohorts to apply the principles. We can only make very limited inferences as to the cause(s) of the observed errors; it was not possible to observe clinicians' reasoning whilst they completed the cases. Image interpretation and anatomical knowledge are commonly cited as reasons for contouring variation (see Chapter 2), but this study highlights other factors such as conceptual understanding and case complexity. Further work exploring the clinical reasoning mechanisms involved in radiotherapy contouring may enable

researchers to understand the mechanisms underlying contouring variation and thereby target interventions.

As with IMRT, the systematic analysis of qualitative errors allows us to target the same principles in future learning exercises (see Chapter 9) and materials (such as a guidance documents i.e. a contouring guideline or atlas). Because qualitative errors vary by case, more extensive sampling is needed to identify contouring errors that could occur within EMBRACE-II and in clinical practice. This does not need to be entirely prospective - common errors can be collated from data from a variety of cohorts (previous educational workshops, other trial quality assurance) provided they contain a representative spectrum of cases and learners or practitioners. In the UK NHS e-Learning for Healthcare programme for image-guided adaptive brachytherapy (Tan, 2010), a list of common scenarios is presented:

Table 6-7 - Common patterns of regression for cervix cancer brachytherapy target contouring

Initial stage (pre-EBRT)	Pattern of regression
IB1	Complete regression
IB2	Complete regression Poor regression
IIB / IIIB “same principles apply”	Complete regression Partial regression No regression
IVA	Good regression No regression

Cases 1 and 2 reflect patterns of partial regression (Case 1) and near-complete regression (Case 2) in stage IIB disease. The e-Learning for Healthcare module states that the target volume principles for stages IIB and IIIB are the same. In the EMBRACE-I trial, 66% of participants had either stage IIB (52%) or IIIB (14%) disease. These data provide some encouragement regarding the representativeness of the accreditation cases, but the contouring assessment still excluded stages 1B1&2 (18% of EMBRACE-I patients), IVA (7%), and IIIB with either complete or no regression (percentage data not available). Further work is required in EMBRACE-II to test the principles of contouring in these scenarios, as well as testing the application of the principles in further cases.

Addressing these errors with practice over multiple high-fidelity simulated cases with manual feedback would be highly resource intensive. If on average it took the two assessors 15 minutes to judge a submission and provide feedback, then for this exercise alone that would add up to

nearly 100 person-hours of work, before starting to count time clinicians spend contouring. As put forward previously, there is a need for a platform to allow efficient testing and teaching over a representative spectrum of clinical scenarios.

A formal analysis of the dosimetric implications of each error would be helpful to prioritise future exercises and guide radiotherapy quality assurance (clinical trials) and peer review (clinical practice). Assessment of clinical impact of individual errors in this study currently relies on expert judgement. Assessment of dosimetric impact of errors, as part of future work, could be aided by the addition of automated 'knowledge-based' radiotherapy planning which would reduce the resources taken and may allow routine application to learning exercises (Lim et al., 2019).

Conformity indices and automated assessment

The data in this study provide further evidence that a uniform JCI cut-off to indicate clinical acceptability is unlikely to be appropriate except in narrow circumstances, such as when a specific cut-off has already been validated. It not only varies by region of interest but also by case, so as to make universal standard setting using this metric alone very challenging, if not impossible. Contours of relatively low conformity may still be clinically acceptable. As noted previously, the *location* of contouring discrepancies and the *clinical significance* of the variation cannot be subjugated to the *degree of conformity* simply because conformity can be quantified.

Brachytherapy organs at risk may be a special circumstance as the most important parameter determining toxicity is the dose to most exposed 2 cubic centimetres of the organ (D2cc). As long as the clinician has contoured the organ in the region of the D2cc the overlap (or lack thereof) with the rest of the reference contour is not critical, although it may be relevant to exploratory analyses.

Brachytherapy targets may also be unique when compared to EBRT. Relatively small variations in the most lateral extent of the HR-CTV contour in the parametrium (e.g. in Case 1) will significantly affect dosimetry for both target coverage and organs at risk, whereas the superior/inferior extent of the HR-CTV contour is unlikely to have significant consequences because in most clinical scenarios a relatively standard pattern of uterine tandem loading (i.e. standard uterine dose distribution) is used.

As in the previous chapter, this discussion is based on analysis using a single planar similarity index and volumetric data, which limits the generalisability of my findings. However, in the recent review of educational interventions for contouring by Cacicedo et al. (Cacicedo et al., 2019), the

most common methods used to assess contours were expert opinion and conformity metrics of planar similarity. A detailed analysis comparing different indices (multiple planar similarity indices, volumetric data coupled with centre of mass, and surface distance) will be conducted as a separate piece of work.

Proponents of conformity index-based assessment may argue that the flaws in manual assessment documented here casts doubt on or even invalidates these findings, and that conformity indices are instead adding valuable objectivity to the assessment process. Certainly, confirmatory studies in from other trial groups and tumour sites would be relatively easy to conduct and may support or refute the generalisability of these findings. More studies analysing the inter-rater variability of contouring assessment would also be interesting. However, despite its flaws, clinical assessment is undoubtedly the gold standard for clinical trial accreditation - potential surrogates such as the JCI need to be tested against it.

Conformity indices are still useful. Some high-fidelity simulations use conformity indices with reference to population statistics as formative feedback to help learners judge their level relative to others. Conformity indices have also been used on a slice-by-slice basis to point clinicians to areas of discrepancy (Conibear, 2018). A standard pre-validated (ROI and/or per case) cut-off may be useful to automatically assess contours with particularly low or high conformity without manual review, saving assessors time if they were to use the case again for further clinicians. Alternatively JCI could be trialled within a compensatory scoring framework where individual measurement error is less important (McKinley and Norcini, 2013). Even then, valuable qualitative feedback to enable clinicians to improve may be lost. An ideal automated assessment (whether formative or summative) would be able to effectively distinguish between competent or non-competent clinicians whilst at the same time providing useful feedback.

6.5 Conclusion

Despite having experience using IGABT target concepts in clinical practice, participating clinicians nevertheless made errors in applying brachytherapy target concepts to standardised test cases with relatively high frequency. As seen in previous studies, the volume to which the radiotherapy dose is prescribed, the HR-CTV, was the most consistently contoured which is reassuring for the impact on patient outcomes. The variability seen in GTV_{res} and IR-CTV (and less frequently organs at risk) contouring could impact the validity of dose-volume analyses in the EMBRACE-II trial if uncorrected. Further quality assurance and education is necessary, and is ongoing.

Conceptual errors carried across cases, but selecting a different case also brought out different errors relating to the anatomy, radiology, or the application of a particular concept. Clinical reasoning studies to investigate the underlying causes of these errors would be valuable.

Conformity index analysis based on planar similarity is not sufficiently accurate to distinguish contours assessed as clinically adequate or inadequate on an individual level in most situations. It is nevertheless helpful to illustrate trends between regions of interest, cases, and cohorts.

Testing (and teaching) across a representative spectrum of cases is warranted - methods to do this in a time efficient manner would be highly valuable as the resources required to provide manual assessment with bespoke feedback limited the validity of the accreditation process.

7 Initial development of Mini-Contour: a low-fidelity radiotherapy contouring simulation

The start of this chapter will briefly review the findings from the ‘analysis and exploration’ phase of this educational design research programme and make the case for the development of a novel low-fidelity radiotherapy contouring simulation. The middle sections document the ‘design and construction’ phase and present the initial specification of a novel low-fidelity radiotherapy contouring simulation - “Mini-Contour” - with the results of early informal user testing. The final sections describe the second iteration of Mini-Contour as evaluated in the usability (Chapter 8) and pilot contouring education (Chapter 9) studies.

7.1 Review of ‘analysis & exploration’ findings

Chapter 2 framed the problem of contouring variation: numerous studies document significant variation between clinicians which can have serious clinical consequences. Improving this situation is challenging especially given variation even between ‘experts’, but errors can be identified along with potential consequences - examples of these were demonstrated in the EMBRACE-II contouring quality assurance presented in Chapters 5 & 6.

Chapters 2, 3, 5 & 6 also highlighted the challenge for contouring assessment, especially with regard to ‘content validity’. Contouring errors vary by case and radiotherapy quality assurance is generally not fully representative of the range of clinical scenarios included in trial eligibility criteria or the variety seen in clinical practice.

Progress in reducing inter-observer variation has been made (especially in the production of contouring guidelines, atlases, and workshops), but overarching educational programmes are less well developed - reported programmes are almost exclusively short term and contouring improvements are often evaluated with conformity indices, which lack assessment validity, and often on same case i.e. there is a lack evidence for contouring skill retention and transfer. Deliberate practice (Chapter 3, Section 3.6) can increase the effectiveness of simulation programmes for practical skills. A high-fidelity approach is instinctive and commonly seen in radiotherapy contouring, but contouring a complete case takes a long time and requires complex software. The potential advantages of low-fidelity simulation can be explained using cognitive load theory (Chapter 3, Section 3.5) and have been borne out by empirical evidence in novice learners in medicine.

Assessment can promote learning and retention (Chapter 3, Sections 3.7& 3.8) and can also give clinicians in established practice insight into their areas for development (Chapter 5) - this can be helpful when methods of helping clinicians adapt to new concepts and change their practice are required. The EMBRACE-II EBRT online contouring programme (Chapter 5) was designed to enable this, but engagement was limited. Reducing the time taken to complete a learning exercise may promote engagement, and focussing on common errors per region of interest and tumour stage might be another way to increase testing and learning efficiency.

Chapter 3 described the powerful effect of feedback for learning clinical skills (Chapter 3 Section 3.8). Given some contouring errors are related to principles or concepts, qualitative feedback may help clinicians to recognise and apply across cases, and correct faulty mental models. A recently presented survey of UK trainees showed a strong desire for qualitative contouring feedback (Evans et al., 2019a).

A low-fidelity simulation could enable innovation in contouring assessment and learning at relatively low software development cost - the flexibility of a smaller code base for low-fidelity software avoids the extensive re-working required to innovate with tools like the Addenbrooke's Contouring Tool (Chapters 5 & 6). Some potential advantages and disadvantages of a low-fidelity software approach for contouring assessment and learning are reviewed in Table 7-1:

Table 7-1 - Potential advantages of low-fidelity software for contouring assessment and teaching

Potential advantages	Potential disadvantages
Lower loading times -> rapid repetitive practice Can include wider variety of cases (clinical variation; assessment content validity)	Time- efficient approach could promote superficial engagement with underlying constructs
Reduce software complexity (cost and cognitive load)	Reduced functional task alignment (e.g. 2D imaging)
Shorter exercises may improve engagement	Learners may perceive software as less representative of clinical practice -> reduced acceptance
Enable deliberate practice targeting specific principles (radiologic interpretation, contouring target concepts, patterns of spread), if feedback incorporated.	May work less well (retention, transfer) for experienced clinicians (expertise reversal effect). Deliberate practice possible (likely slower-paced) with high-fidelity software.

These potential advantages merit further evaluation. Developing a low-fidelity contouring tool does not preclude applying lessons learned to a high-fidelity tool; such a simulation would have to be evaluated against high-fidelity simulation (as the current “standard of care” in radiotherapy education) in future.

7.2 Design & construction

7.2.1 Development framework

The framework proposed by Olszewski & Wobrink for the development of serious games and virtual simulation in medical education (Olszewski and Wolbrink, 2017) formed a structure for the initial development process:

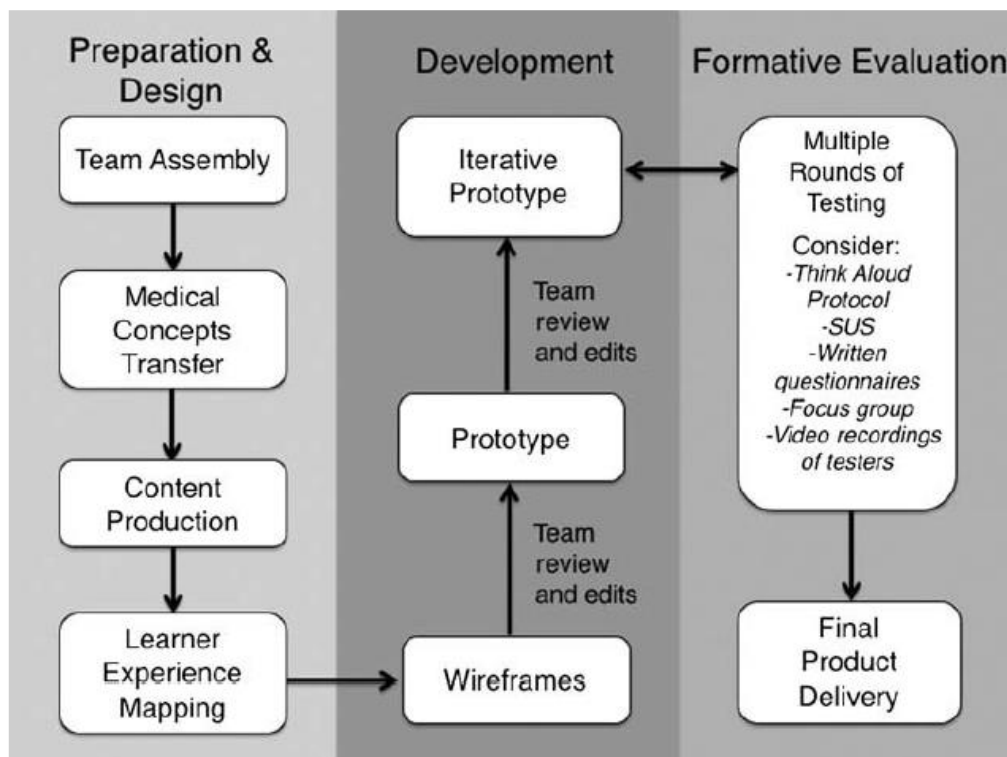


Figure 7-1 - Olszewski & Wolbrink's framework for virtual simulation development. Reproduced from Olszewski & Wolbrink, 2017

7.3 Foundations, initial specification & early testing

In 2014, a simple flash-based online contouring simulation was developed by Dr Tan. This was a greatly simplified two-dimensional representation of the TVD process, but allowed the learner to attempt interactive exercises and receive rapid visual feedback in the form of a solution contour:



Figure 7-2 - Low-fidelity contouring simulation developed by Dr Tan in 2014

The flash-based contouring tool formed the initial ‘wireframe’ for the design, but could not store user data, and so lacked the necessary functionality to enable analysis of learner activity and performance. In addition, we were conscious that flash-based technologies would not be supported by web browsers from 2020.

The initial aim was to develop an low-fidelity simulation of radiotherapy contouring with the following functionality:

- A simple interface with contouring on one or more image slices in one plane (i.e. two-dimensional); contouring to be feasible using a laptop trackpad
- Access to an exercise within 20 seconds and completion of a single delineation exercise in less than 3-5 minutes when connecting over an NHS internet connection, to allow rapid practice
- Storage of participant contour data in a database for analysis, including date and time information to allow analysis of participant time use
- After participant contour submission, obtain feedback by visualising one or more reference contours superimposed over the participant’s contour
- Ability for a teacher to superimpose and display all learner contours for group discussion
- Creation of a new case from a teacher interface

The hypotheses for development of this simulation were:

- A simplified online radiotherapy delineation simulation will enable testing of specific principles or aspects of delineation in less than 3-5 minutes
- This simulation will reproduce errors seen in high-fidelity contouring simulation

7.3.1 Development – initial phase

A web developer – Mr Adam Dorling – was identified through personal contacts. In November 2017 myself, Dr Tan (PhD supervisor) and Mr Dorling met at the ‘team assembly’ and ‘concept transfer’ stage to discuss radiotherapy contouring, our pedagogical approach and the software requirements and possibilities.

I then drew up a one-page specification outlining the functionality required (see Appendix Section A.7.1) for the initial prototype. This functionality was successfully delivered on a budget of £250 and within a 4 week timescale, for which I am incredibly grateful to Mr Dorling. Qualitative feedback in addition to a ‘solution’ contour (in feedback terminology this is ‘knowledge of the correct response - see Section 3.8) was discussed for the initial prototype, but because the required functionality was more complex to design this was planned for the second iteration.

Figure 7-3 to Figure 7-5 below show the user workflow through a learning exercise:

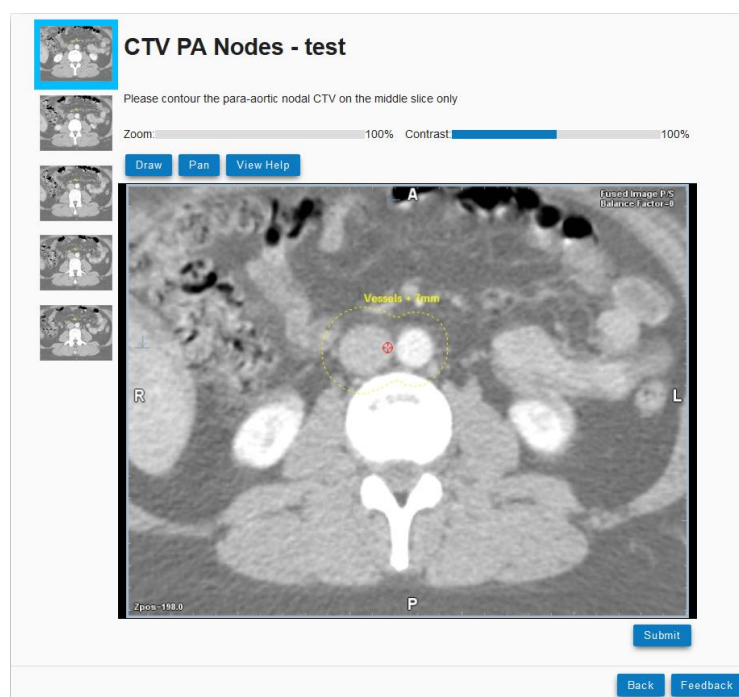


Figure 7-3 - Mini-Contour prototype: user view on accessing a learning exercise. You can see that a guide (dotted yellow line) is provided as a marker to indicate the standard (7mm) expansion from the major vessels (aorta and inferior vena cava).

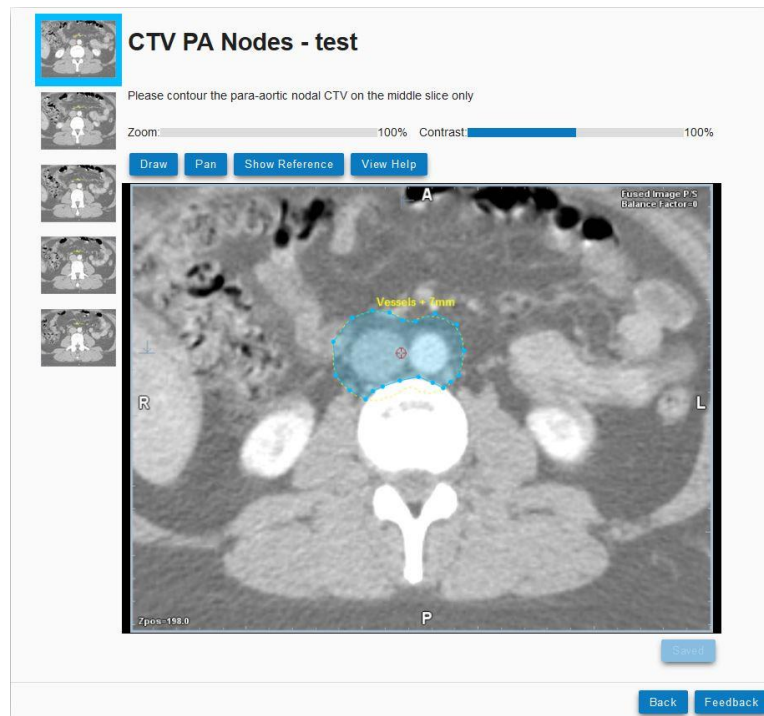


Figure 7-4 - Mini-Contour prototype: the user contours the target and presses the submit button. Here the user has avoided contouring the bone - this is correct as there is no infiltration of the cancer cells in to the bone. The 'Show Reference' button appears, which the user can then click

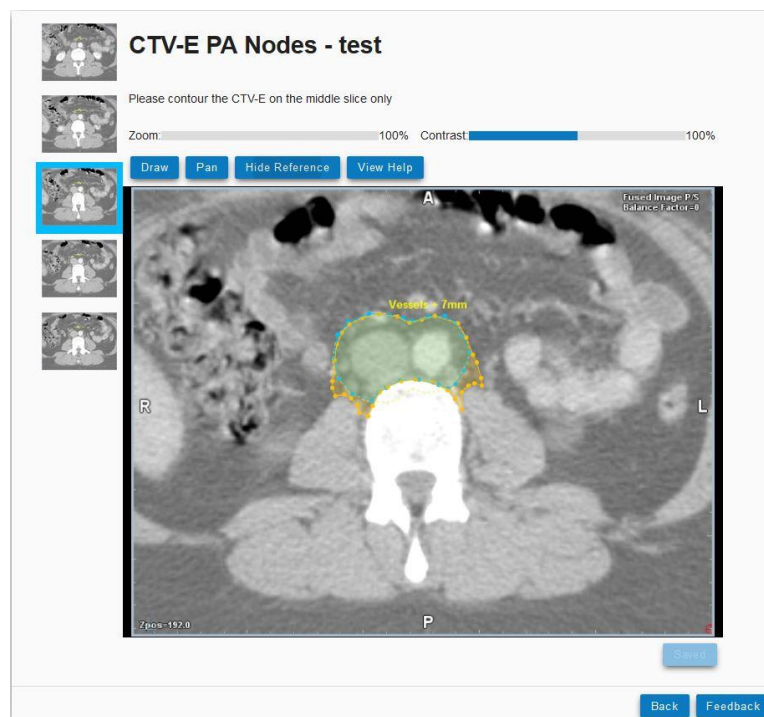


Figure 7-5 - Mini-Contour prototype: the reference contour is revealed. Here the participant has missed the left lateral area of the target volume; a lymph node region where the cancer may recur which should be included

7.3.2 Initial design and early user testing

During the following 3 months (early January – March 2018) early learner experience mapping and preliminary testing was conducted, which exposed some technical problems and identified areas for future development. Whilst this testing was not systematic, informal usability testing can nevertheless inform development (Sandars & Lafferty, 2010) and enable rapid development cycles. This was conducted via:

- Individual interviews with oncology consultants, trainees and physicists
- An anatomy lecture with 170 medical students
- The 2018 annual EMBRACE group meeting with >40 oncology consultants and a live workshop featuring 6 delineation scenarios

Individual interviews

These sessions consisted of me sitting next to or video-conferencing with an individual for around 15 to 30 minutes. Whilst the user registered on the software and navigated through an example case, they tried to articulate their thoughts on the process whilst I made brief notes. Users included oncologists from Cambridge (consultants and trainees), Norwich (consultant) and Vienna (consultants and physicists).

The interviews highlighted a major technical issue – the software was developed for use on the Google Chrome browser, but some users did not have this on their computers. Use of a different browser such as Firefox or Internet Explorer led to incorrect display of the software. This was a significant barrier to use and was addressed with high priority as an interim modification.

Many users did not read the instructions (“View Help”) prior to starting to attempt contouring. This meant that some held down the mouse and then moved the cursor to ‘draw’ (as seen in Microsoft Paint and Adobe Photoshop) rather than clicking to make a contour point. The case instructions (seen above the zoom bar) were often missed by the users. It was also not clear to the users that the ‘Feedback’ button referred to them sending their impressions of, or problems using the simulation – they often inferred that this button would result in them receiving feedback on their performance.

The simulation was set up for 2D contouring on a single image set, but users felt that in some situations context from other image series was essential - either a related plane (e.g. the sagittal

plane when viewing axial images) or another timepoint (e.g. the diagnostic MRI scan when looking at an MRI of the cervix tumour during brachytherapy treatment).

A common point made about the usefulness of the tool for learning was that some qualitative feedback was desired – it was felt that an *explanation* of the discrepancy between the user was important to maximise learning.

Medical student anatomy lecture

Two case scenarios were presented to students in a 15-minute session at the end of an anatomy revision lecture. One scenario asked students to delineate the kidney on axial CT images and the other scenario asked students to delineate the portal vein on a picture of an anatomical prosection. Unstructured written feedback was collected via a smartphone-based system.

Over 170 students registered and commenced delineation simultaneously, which lead to a large server load and some delays – the majority of students took 20-30 seconds before being able to start contouring. Both exercises were then completed successfully.

Feedback received was generally positive. It highlighted some usability issues - students reported that they would like to be able to delete their own contour and start again, and be able to see an ‘answer’ contour. This showed that these functionalities (which were already present) were not sufficiently obvious or intuitive.

EMBRACE group 2018 annual meeting

6 case scenarios were presented, worked through and discussed during a 1 hour 20 minute delineation workshop involving more than 50 participants at the EMBRACE group 2018 annual meeting. The audience contained a mixture of clinicians – most were from centres that had been accredited to the EMBRACE-II trial, however some were from centres that were undergoing the accreditation process and some outside of it completely.

3 EBRT and 3 brachytherapy exercises were trialled. After each exercise, the results of the delineation were displayed on the main projector and discussed amongst the workshop facilitators and audience.

Despite clinicians delineating on only one axial image, the exercises revealed that common errors seen in the radiotherapy quality assurance accreditation exercises were reproduced:

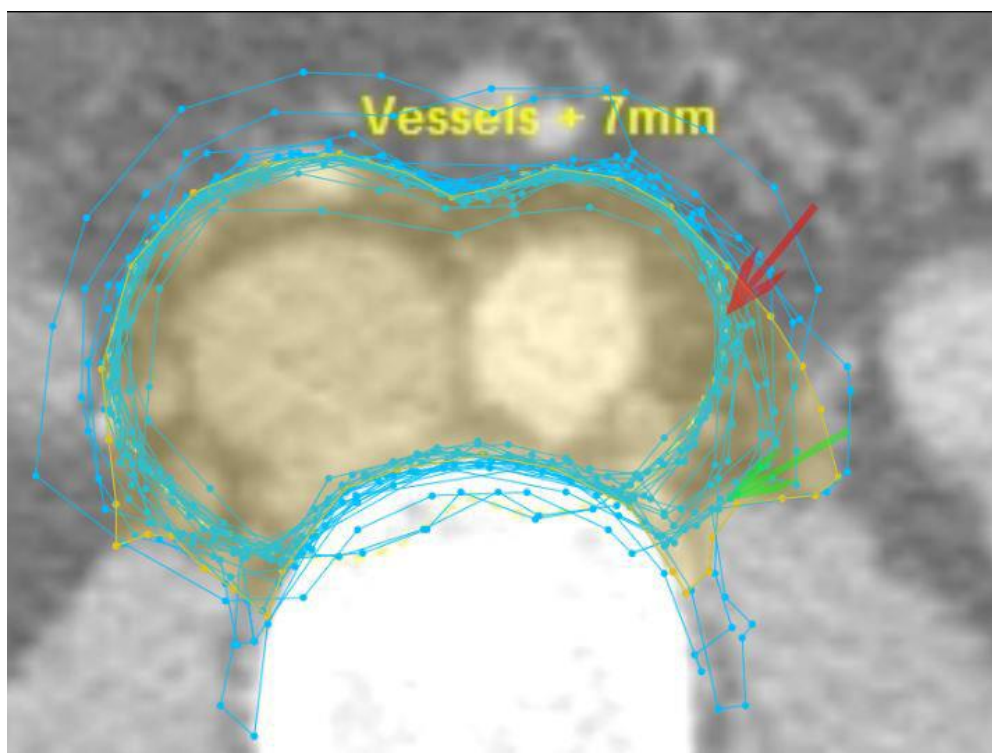


Figure 7-6 - EMBRACE 2018 workshop: participant contours from an external beam radiotherapy exercise. Many participants contoured around the 7mm expansion (red arrow), rather than extending the contour laterally to reach the psoas muscle.

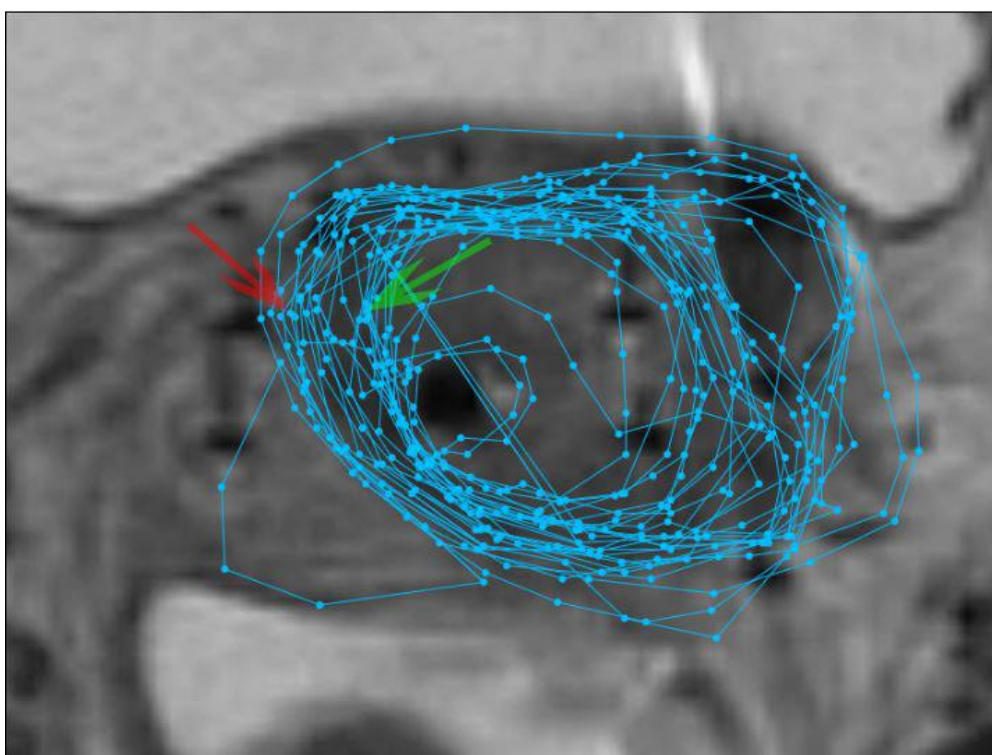


Figure 7-7 - EMBRACE 2018 workshop: delineation of the residual gross tumour volume (GTVres) at time of brachytherapy. Delineation should extend to the edge of the high signal region on MRI (green arrow). Many participants extended the delineation beyond this region (red arrow).

Other issues seen across cohorts

Users contouring on the wrong image slice was common. This was rectified by changing the CT images uploaded (superimposing “contour this slice” on the relevant image slice, and adding an instructions image as the first “image slice”) rather than a software redesign.

As the number of exercises and courses increased, course management and exercise selection became increasingly complicated - it was difficult for users to “see the wood for the trees” when they had access to all exercises. It was clear that functionality to manage users access to and display of the relevant exercises would be needed.

7.3.3 Summary of findings from initial testing

The initial testing showed the ability of informal testing to highlight important usability issues (‘delete contour’ functionality not sufficiently intuitive, case instructions often missed) and technical issues (browser incompatibility, high server load for large groups).

The workshop testing demonstrated the feasibility of using this tool in live workshops to test multiple different contouring scenarios. It has also shown that brief low-fidelity exercises can identify similar errors as those seen in high-fidelity simulation, for example e.g. extension of para-aortic lymph node delineation to the left psoas muscle (Figure 7-6).

The testing also demonstrated the feasibility of applying this tool to other settings such as anatomy teaching. Testing radiological interpretation - for example identifying bony fractures, cancer metastases, or pulmonary emboli - would be another possible use for this type of simulation.

In highlighting the absence of qualitative feedback provided to the user, this testing also identified a key opportunity which we sought to address with subsequent development.

7.4 Next iteration

The principal aims of the next development cycle were to:

- Enable qualitative feedback
- Improve learning exercise management
- Improve the interface to make case instructions and contouring functionality more intuitive

Enabling qualitative feedback - the 'learning zone' concept

Given the lack of consensus in delineation, there is no single 'correct' answer. However, there may be expert consensus on 'incorrect' answers for a particular tumour type, protocol and case. Errors, such as those seen in the EMBRACE-II quality assurance (Chapters 5 & 6), fall into two categories:

- Tumour-related targets that are incorrectly excluded from or incompletely included in the user's contour ("under-contouring")
- Organs at risk or other normal tissues not at risk of tumour spread that are incorrectly included in the user's contour ("over-contouring")

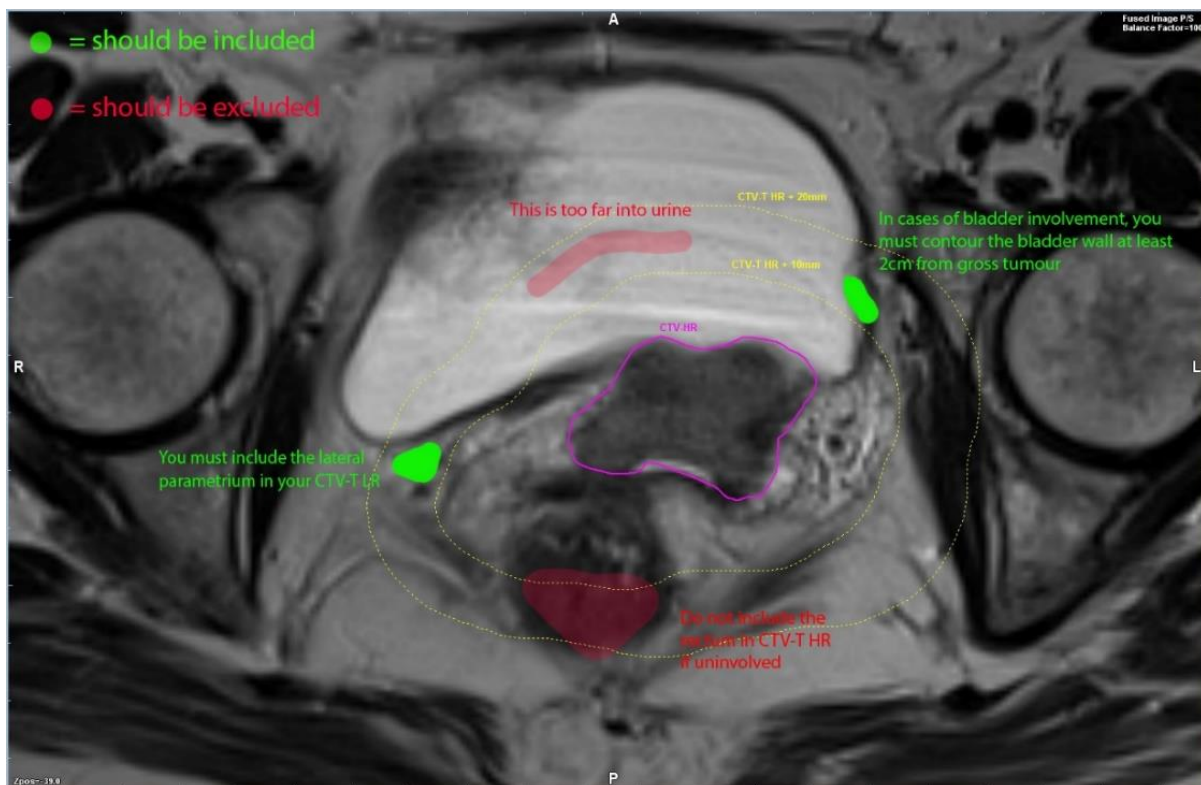


Figure 7-8 - First sketch of the 'learning zone' concept to allow automated qualitative feedback. The task in this case is to expand the EBRT CTV-High Risk (purple contour) to a "low risk" clinical target volume (CTV-LR_{init}). Green areas (tumour-related targets) should be excluded from the contour, and red areas (normal tissues) should be included.

Enabling qualitative feedback for specific areas is technologically relatively simple: an automatic test for overlap can be computed between the user contour and the defined learning zones. The red areas (Figure 7-8) should be completely *excluded* from the contour. If the user contour overlaps with any part of a red area, the user then can receive an 'error' message with associated feedback (see also Figure 7-9.E below). Conversely, if the user successfully avoids this area then

they can receive a congratulatory message. The exact converse holds true for green areas, which would need to be entirely *included* for the user to receive a congratulatory message.

Other development

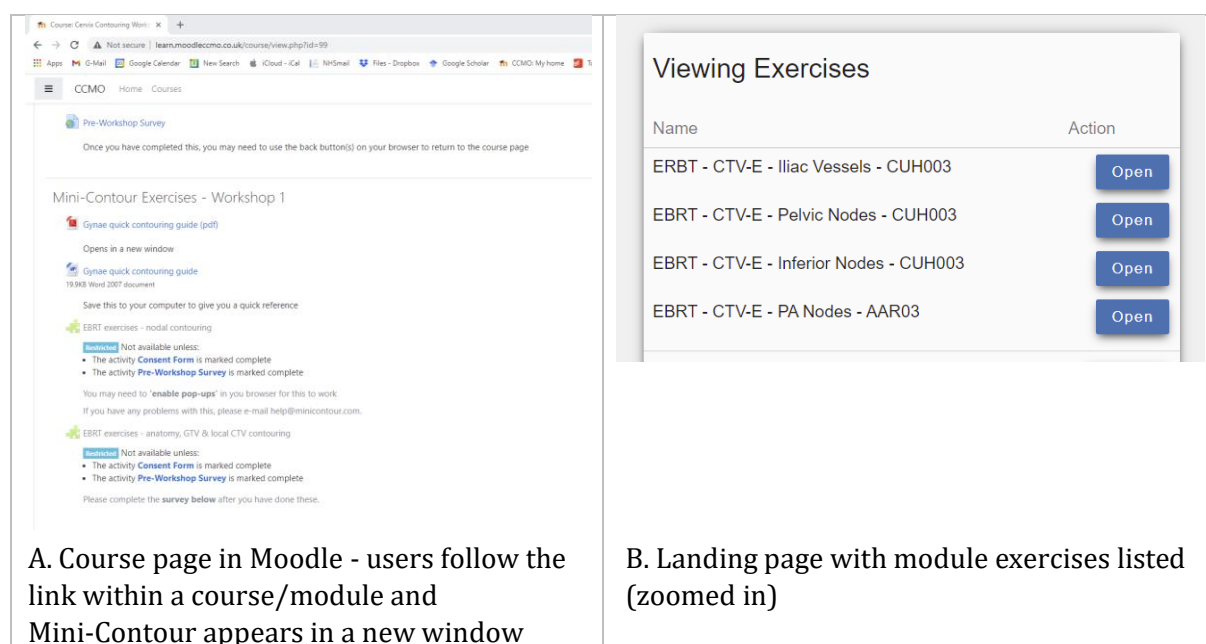
Rather than develop a new exercise or course management system, we decided to use those already developed within learning management systems (such as Moodle®) - we enabled access to Mini-Contour from within specific modules using Learning Tool Interoperability (LTI) standards (IMS Global, 2020) which pass login and course data between educational systems securely.

Functionality was added to record the time taken at each step of the contouring process (logging in, loading data, contouring, reviewing feedback). This was in order to test our hypothesis that completing an exercise would be possible within 3-5 minutes, and to explore user flow through a series of exercises.

For development of the new interface 'wireframes' were created in Microsoft PowerPoint.

7.4.1 Mini-Contour version 1.0

Figure 7-9 below illustrates the user interface and flow through a sample exercise. The version of Mini-Contour shown (v1.0) is the software evaluated for formal usability testing in Chapter 8 and feasibility and learning outcomes within educational workshops in Chapter 9.



EBRT - CTV-E - Pelvic Nodes - CUH003
Clinical information on the first slice. Please contour the marked CT slice only

Clinical Information

35 year old lady
Figo stage IIIA Cervix Cancer -> For chemoradiotherapy
45Gy / 25# over 5 weeks

Instructions: contour the Elective Nodal CTV on the marked slice only*

Tip – include the iliac and pre-sacral nodes

* marked "CONTOUR THIS SLICE"

Draw

Draw Tools

Draw **Pan**

Zoom: 100%

Contrast: 100%

Contour

Your contour

Add Another Contour

Comment **Submit**

Back **Mini-Contour Help**

C. Initial view of the exercise with instructions on the first image slice

EBRT - CTV-E - Pelvic Nodes - CUH003
Clinical information on the first slice. Please contour the marked CT slice only

Clinical Information

35 year old lady
Figo stage IIIA Cervix Cancer -> For chemoradiotherapy
45Gy / 25# over 5 weeks

Instructions: contour the Elective Nodal CTV on the marked slice only*

Tip – include the iliac and pre-sacral nodes

* marked "CONTOUR THIS SLICE"

Draw

Draw Tools

Draw **Pan**

Zoom: 100%

Contrast: 100%

Contour

Your contour

Add Another Contour

Comment **Submit**

Back **Mini-Contour Help**

D. User view of the relevant CT slice for contouring. The user contour is in blue. Dotted yellow lines (very small in this view) show a pre-provided 7mm expansion around the blood vessels. A ruler denoting 10mm (also very small without zooming) is also provided.

EBRT - CTV-E - Pelvic Nodes - CUH003
Clinical information on the first slice. Please contour the marked CT slice only

Draw **Contours**

✗ You should exclude Psoas Muscle (L)
The psoas muscles (the left psoas is contoured here) should be excluded
[Comment / Disagree](#)

✓ You included Extend to Ant. Sacrum
The contour should be extended to the region between the vessels and the anterior sacrum
[Comment / Disagree](#)

✓ You included Pre-Sacral Nodes
The pre-sacral nodes around this region. Do not exclude bowel.
[Comment / Disagree](#)

[Back](#) [Mini-Contour Help](#)

E. The feedback view is revealed after the user presses 'Submit'. Solution contours are shown in yellow with the user contour in blue.

EBRT - CTV-E - Pelvic Nodes - CUH003
Clinical information on the first slice. Please contour the marked CT slice only

Draw **Contours**

✗ You should exclude Psoas Muscle (L)
The psoas muscles (the left psoas is contoured here) should be excluded
[Comment / Disagree](#)

✓ You included Extend to Ant. Sacrum
The contour should be extended to the region between the vessels and the anterior sacrum
[Comment / Disagree](#)

✓ You included Pre-Sacral Nodes
The pre-sacral nodes around this region. Do not exclude bowel.
[Comment / Disagree](#)

[Back](#) [Mini-Contour Help](#)

F. Learning zone feedback is displayed on the right hand side - hovering over the relevant box with the mouse highlights the corresponding learning zone on the image (below)

Figure 7-9 - User flow through a learning exercise in Mini-Contour version 1.0

8 Mini-Contour: usability study

8.1 Introduction

Educational software has the potential to enable users to effectively and efficiently accomplish their learning goals, but the accessible, engaging and effective tool envisaged in the mind of the designer must translate to such an experience for the user. Rubin & Chisnell describe software as truly usable when: *“the user can do what he or she wants to do the way he or she expects to be able to use it, without hindrance, hesitation, or questions”* (Rubin and Chisnell, 2008, p.4). The International Organization for Standardization (ISO) have defined usability as: *“The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”* (ISO, 2010). Usability testing is an important component of educational design research projects (Chen and Reeves, 2020).

The aim of formative usability testing is to: *“inform product development by the early identification and rectification of usability problems early during product development”* (Sandars and Lafferty, 2010, p.960). Rubin and Chisnell’s outline of the three core questions about product usability and appropriate study methods to answer them (Rubin and Chisnell, 2008) are displayed in Figure 8-1. Prototyping, walk-throughs and elements of participatory design were included in the early phases of development reported in Chapter 7. Other methods such as heuristic evaluation - a process whereby an expert in usability and/or human factors research reviews the product - can precede or complement user testing. However, a key aspect of formal usability testing is the emphasis on evaluation by the “end user” rather than just the experts - *“No amount of review, assessment, validation, or other metric conducted by experts can confirm the usability of a course. Only the user can do that”* (Barnum, 2008).

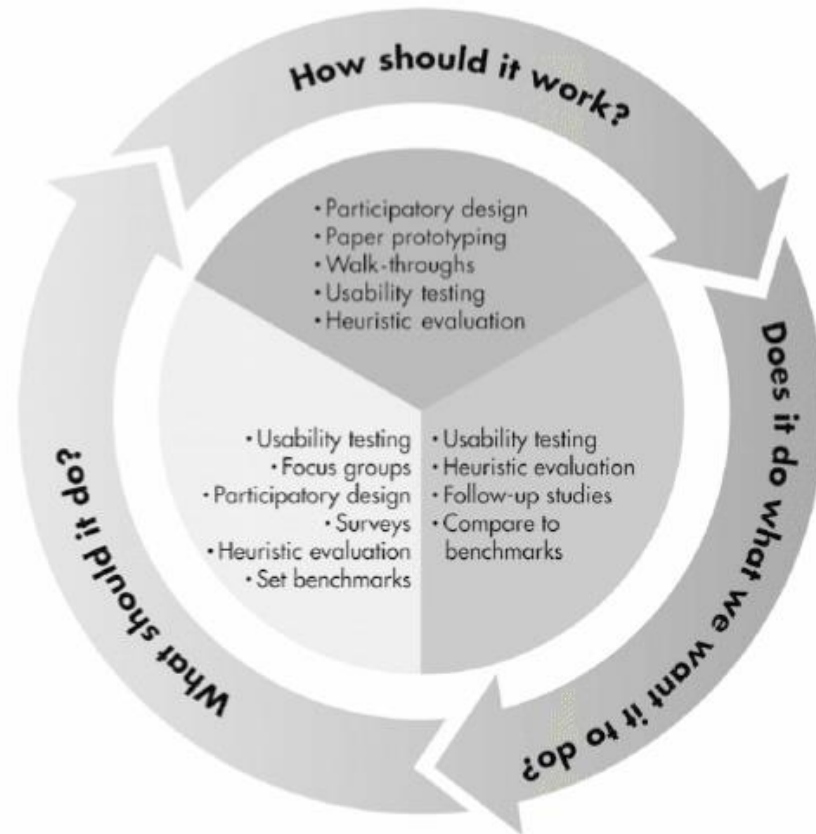


Figure 8-1 - Usability questions and methods to address them. Reproduced from Rubin & Chisnell (2008, p.15).

Studying the user interacting with the tool ‘live’ is essential to understanding the user perspective (Sandars and Lafferty, 2010). This chapter therefore follows Chapter 7 with a more rigorous and detailed evaluation as part of the second iteration of Mini-Contour development. The studies in Chapter 9 report aspects of Mini-Contour usability (for the same version) at a larger scale but in less depth.

8.1.1 Aims

The primary purpose of this study was to explore the usability of the Mini-Contour tool, and to identify software or design issues that impair the user experience. Study endpoints are listed in Table 8-1 below. This study also aimed to:

- Explore users’ expectations and acceptance of the tool, and particularly the automated feedback
- Garner users’ ideas for future development

- Collect preliminary observations of clinicians actively contouring to provide initial qualitative data on their cognitive processes

Table 8-1 - Usability study endpoints

Type	Domain	Data collection & analysis
Primary	Usability problems	Issues coded by severity and frequency Successful completion rate, by sub-task
Secondary	Users' impressions and acceptance of the technology	Practical session comments and interview sessions - thematic analysis. Unified theory of acceptance and use of technology (UTAUT) questionnaire
	Perceived ease of use, usefulness, enjoyment and satisfaction	UTAUT questionnaire, subsequent discussion & semi-structured interview - thematic analysis
	Perceptions of automated feedback	Practical session comments and interview sessions - thematic analysis
	Perceived fidelity	Post-task questionnaire and semi-structured interview - thematic analysis
	Users' suggestions for improvement	Practical session comments and interview sessions - thematic analysis

Although the above endpoints guided analysis, the investigators were alert to unexpected insights emerging from the data, as part of the inductive process of qualitative data analysis.

8.2 Methods and materials

8.2.1 Participant recruitment

Nielsen and Landauer (Nielsen and Landauer, 1993) suggest that the best method of conducting usability testing is to run multiple tests with small groups, as the goal should be to improve design and not just document weakness. Their modelling indicates that a sample size of 5 users is sufficient to pick up 85% of the usability problems. Later authors have however stressed the importance of context and appropriate sampling in defining the numbers to be studied (Lewis, 2014).

Inclusion criteria

Participants were required to:

- Be fluent in English
- Have access to a computer and internet access, or be able to travel to Cambridge for the study
- Be a qualified or trainee Clinical/Radiation Oncologist
- Have no previous experience of using the Mini-Contour software

Recruitment process & timeline

Recruitment was initiated through an e-mail to training and professional networks within the South East of England (Appendix A.8.2). Direct approach by the study investigators was not permitted. The e-mail contained a brief introduction, practical arrangements, and a link to the participant information sheet.

Participants took part in their own time and so were offered a nominal £20 payment in appreciation. When a person expressed interest in participating, an investigator (SLD) contacted them to explain the study further, answer any questions, and arrange a convenient time to conduct the study if the potential participant was willing to proceed.

The initial recruitment e-mail was sent out in January 2019. Data collection ran from February 2019 to May 2019.

8.2.2 Study procedures

Study procedures are shown in Figure 8-2. Interviews could be conducted over web-conference using the participant's own computer, or in-person using a standardised setup. All interviews were moderated by a single facilitator (SLD). The web-conferencing software [Zoom] allowed capture of both the participant's screen and their facial expressions, if they had a web-cam.

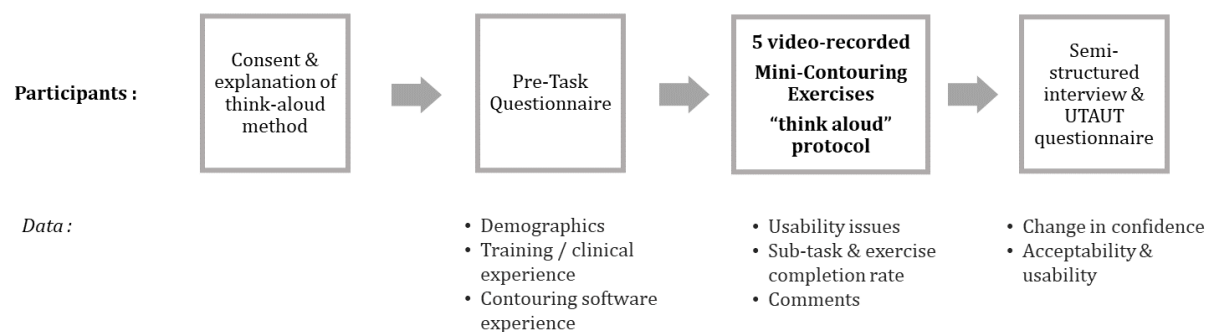


Figure 8-2 - Usability study procedures flowchart

After completing consent process, the facilitator explained the think aloud process using a structured “test script” (see Appendix A.8.3 (U.S. General Services Administration Technology Transformation Services, Barnum, 2011)). Recording started before the participant commenced the web-based pre-task questionnaire (see Appendix A.8.4) so that any relevant spoken comments could be captured.

Participants were then provided with two sets of instructions: a 3-page user-guide for Mini-Contour, and brief (1-page) contouring guidance. A contouring atlas (i.e. pictures of worked example(s)) was **not** provided, as I was interested to see whether users would navigate to one of their own accord.

Participants then worked through 5 exercises and were asked to “think aloud” as they went through - an established method for usability testing (Ericsson and Simon, 1980, Lewis, 1982, Yen and Bakken, 2012). The test script outlined the importance of users vocalising their thoughts as they worked through and they were reminded regularly during the exercises if needed.

Exercises consisted of a tutorial exercise (contouring the mandible - a simple task for almost all radiation oncologists), after which they could choose between 4 ‘head and neck cancer’ or 4 cervix cancer exercises. A deliberate mistake in one of the learning zones was included in the final case (‘include’ switched to ‘exclude’), to see how users would react.

After completing the exercises, participants worked through a post-task questionnaire and semi-structured interview (Appendix A.8.5). The post-task questionnaire was derived from the unified theory of acceptance and use of technology (UTAUT) questionnaire - a 31-item survey derived from a theory-based model which predicts around 70% of variance in user intentions to use information technology (Venkatesh et al., 2003). Eight Open-ended questions then explored users’ perceptions of simulation fidelity, general experience, experience of automated feedback, and suggestions for improvement.

8.2.3 Data analysis

Recordings were transcribed by a single investigator (SLD) - pairing speech with the associated user actions - see Figure 8-3 for an example. The time taken for this process was recorded.

Time	Speaker	Transcript	Actions	Usability issue Severity
03:14	Participant_1	So, I'm gonna go minicontour user guide ... but probably not spend too much ... because ... technical.	Goes into user guide	
03:25	Participant_1	Ok ... we'll see if my computer works ... this will be more obvious when I open it. So what I might do is leave that open and start	Scans first page of user guide (~20 seconds)	
03:56	Participant_1	Ok so there's a mini-contour user guide and there's a Head & Neck nodes user guide. So I'll just see what this says .. If it's any different or quite generic .. Taking a little bit of time ... OK so this is about protocols and things .. Oh gosh ... fine ... OK, again, something to keep for reference	Opens head and neck contouring instructions (open for about 20 seconds)	
04:18	Participant_1	Ok ... mini-contour exercises .. Much easier to see what I'm doing ... What I don't know once I've started it	Follows link to Mini-Contour	
04:30	Participant_1	Um ... mandible sounds easy I'll start with mandible	Chooses 'Mandible' exercise, entering contouring interface	
04:32	Participant_1	Clinical information: oropharyngeal. Treated with chemoradiotherapy. Contour the mandible on the middle slice only .. Ok, so that's 3 down.	Reads the instructions slide Goes to slide for contouring & reads the text there	
04:50	Participant_1	Ok ... I think I'd know ... bone is not too bad.	Starts making a point	
04:55	Participant_1	Oh, it's not a drag one	Making a line and connecting dots	2

Figure 8-3 - Example of transcribed speech and user actions for the Mini-Contour usability study

After transcription, data were analysed in a 5-stage process as described by Wellington (Wellington, 2015, p.260-264). Coding was performed by a single investigator (SLD) without computer-assisted qualitative data analysis software (such as NVivo®), which although powerful can interfere with the researcher's immersion in the data, especially if they have relatively little qualitative research experience (University of Nottingham, 2018).

Table 8-2- Stages of qualitative data analysis. Adapted from Wellington (2015) p. 260-4

- 1) Immersion in the data (transcription, re-viewing, highlighting)
 - 2) Reflecting
 - 3) Analysing
 - a. Dividing data into units of meaning
 - b. Filtering units that can be used
 - c. Categorising or 'coding' the units of meaning
 - d. Merging similar units into a single theme
 - 4) Recombining/synthesizing data*
 - 5) Looking for linkages, contrasts and comparisons between the categories *
- * also known as 'constant comparison'

Users' actions whilst working through the Mini-Contour exercises were categorised into pre-specified activities:

- Navigation to/from an exercise(s)
- Accessing exercise content (including viewing images)
- Drawing a contour
- Editing a contour
- Submitting a contour

- Reviewing feedback and reference contour(s)

Usability issues were coded for severity as per Rubin & Chisnell's schema ((Rubin and Chisnell, 2008) - see Table 8-3) and tabulated with frequency counts. Comments or actions unrelated to usability issues were analysed thematically (Braun and Clarke, 2006).

Table 8-3 - Severity grading of usability issues. Reproduced from: Rubin & Chisnell. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests (p.262-263), Wiley 2008

Severity Rating	Severity Description	Severity Definition
4	Unusable	The user either is not able to or will not want to use a particular part of the product because of the way that the product has been designed and implemented.
3	Severe	The user will probably use or attempt to use the product but will be severely limited in his or her ability to do so. The user will have great difficulty in working around the problem.
2	Moderate	The user will be able to use the product in most cases, but will have to take some moderate effort in getting around the problem
1	Irritant	The problem occurs only intermittently, can be circumvented easily, or is dependent on a standard that is outside the product's boundaries. Could also be a cosmetic problem.

Quotation formatting

Table 8-4 details the punctuation marks denoting how participants' verbatim quotations have been transcribed and/or presented:

Table 8-4 - Punctuation denoting features or framing of participant quotes

Punctuation	Indicates:
"... Quote"	The beginning of a transcribed sentence or phrase has been omitted
..	Participants' natural pauses in speech
[...]	A section of a quotation has been omitted
[user action]	Describing a participant action performed at the same time as the quotation
"Quote ..."	The end of a transcribed sentence or phrase has been omitted

8.2.4 Ethical approval

This study was approved by the University of Nottingham Faculty of Medicine & Health Sciences ethics committee. It was exempt from national NHS & Health Education England approval. Participants gave written informed consent digitally.

8.3 Results

8.3.1 Participant demographics

Five participants were recruited. All were Clinical Oncologists (Table 8-5): one consultant (C_03) and four trainees, two of whom had three or more years of specialist oncology experience and are identified, to aid description, as 'senior trainees' (ST_02 and ST_04) and two of which had less than three years' experience and are identified as 'junior trainees' (JT_01 and JT_05).

Four participated via web-conference and one live. All five chose to complete the head and neck cancer exercises, which consisted of a tutorial (mandible), one organ at risk (parotid) and three elective lymph node CTV contouring (levels 2,3 & 4 (Grégoire et al., 2014b, Grégoire et al., 2014a)). Both of the senior trainees had clinical experience of head and neck cancer treatment, including radiotherapy contouring, whereas neither of the junior trainees did. The consultant was recently qualified but not a site-specialist in head and neck cancer.

Interviews ranged from 32-57 minutes (mean 46) and the transcription of the participants' speech and paired actions took an average of 68 minutes per 10 minutes of recording, i.e. approximately a 7:1 ratio with an average of just over 5 hours of transcription per user.

Table 8-5 - Usability study participant characteristics

Code	Level; Time in role	Previous simulation experience	Head & Neck cancer experience	Computer
JT_01	Trainee; 1 year	None	None (also no gynae experience)	NHS PC
ST_02	Trainee; 3 years	Low-fidelity online tool [CCMO - see Chapter 7]	6 months ~ 1 year previously	Mac Laptop
C_03	Consultant; < 1 year	High-fidelity online tool - ESTRO FALCON programme RTQA	6 months ~ 5 years previously	NHS PC
ST_04	Trainee; 4 years	None	12 months ~ 2 years previously	Mac Laptop
JT_05	Trainee; 2 years	None	None (also no gynae experience)	Windows Laptop

In the pre-task questionnaire the participant with experience of a high-fidelity contouring simulation (C_03) noted: “FALCON tools were slow to get to grips with, which was an obstacle to learning the material”.

8.3.2 Usability issues by contouring activity

Usability issues coded by activity are illustrated in Figure 8-4 below, and listed in full in Appendix Table A.8-1.

The exercise completion rate was 100% and no ‘critical’ usability issues were identified.

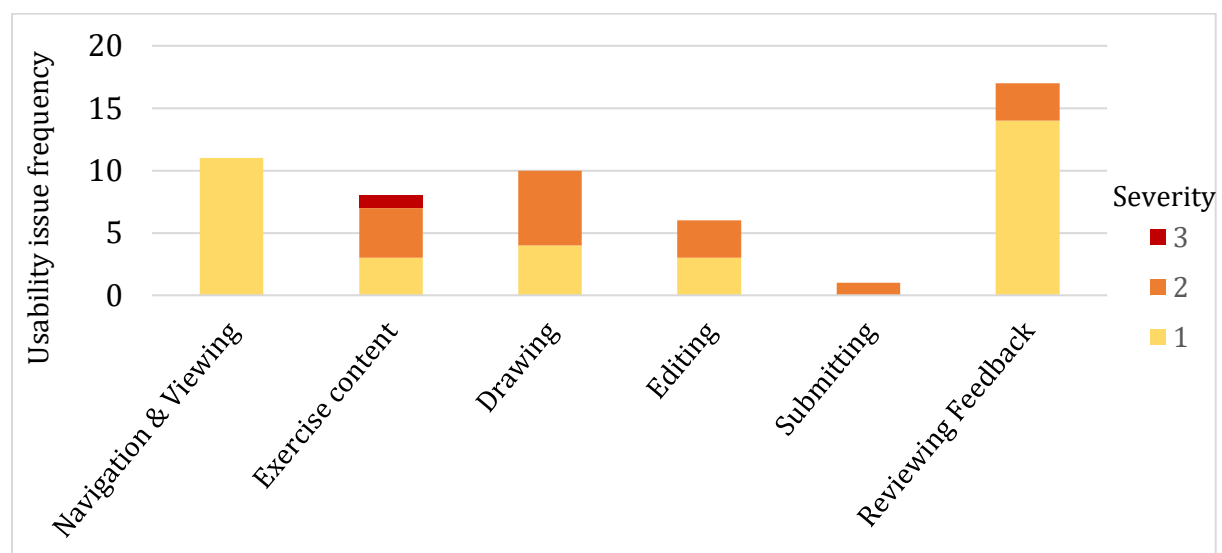


Figure 8-4 - Usability issue frequency coded by activity

As the exercises proceeded, usability issues decreased in frequency as users became more familiar with the interface (Figure 8-5.A).

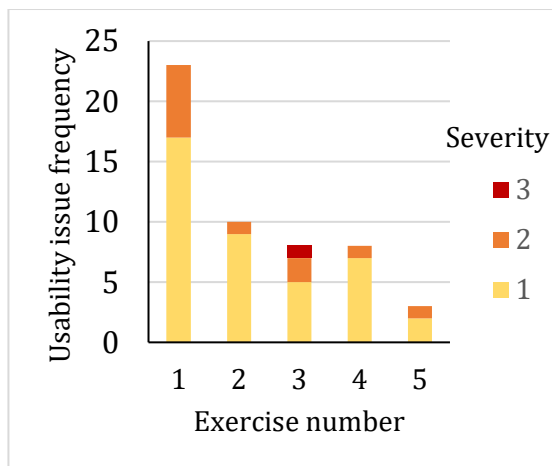


Figure 8-5 - Usability issues by exercise number, graded by severity.

Navigation and viewing

Navigation was generally straightforward. Users generally expected or hoped that Mini-Contour would store their previous attempts for display, and that the list of exercises would indicate previous attempts - JT_05: “Now, I would find it very satisfying if they had some ‘tick’ saying you've attempted and you've failed. ... [...] ... I would find that motivational to get it all done, so all the lights are green”.

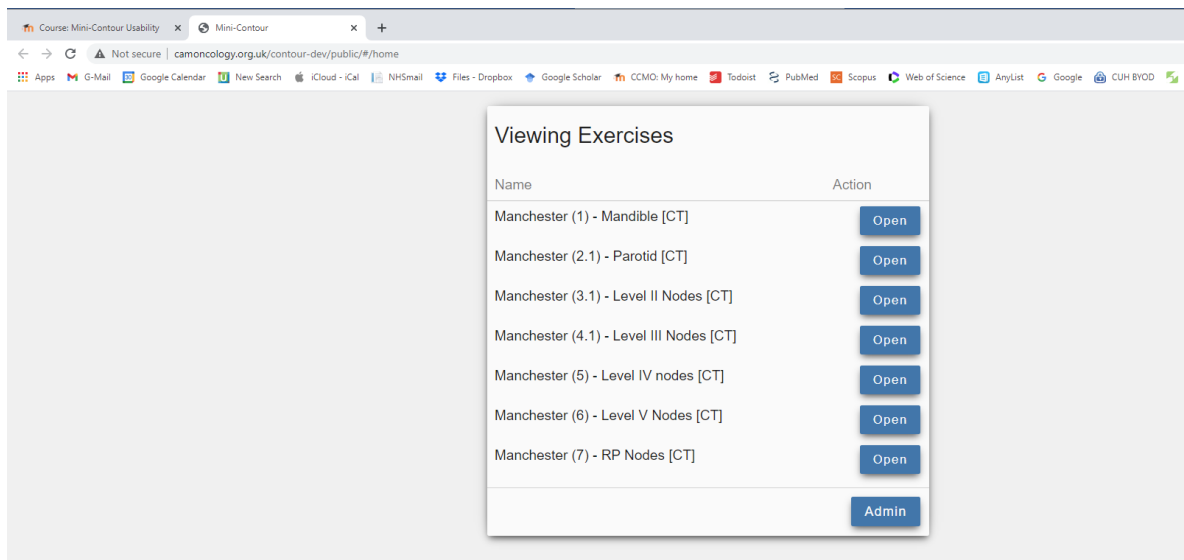


Figure 8-6 - Mini-Contour exercise menu page. Users expected or hoped that the tool would track and display their previous attempts.

For two users the interface did not fill their screen, either due to browser zoom settings or because they made space for the contouring guidelines (Figure 8-7):

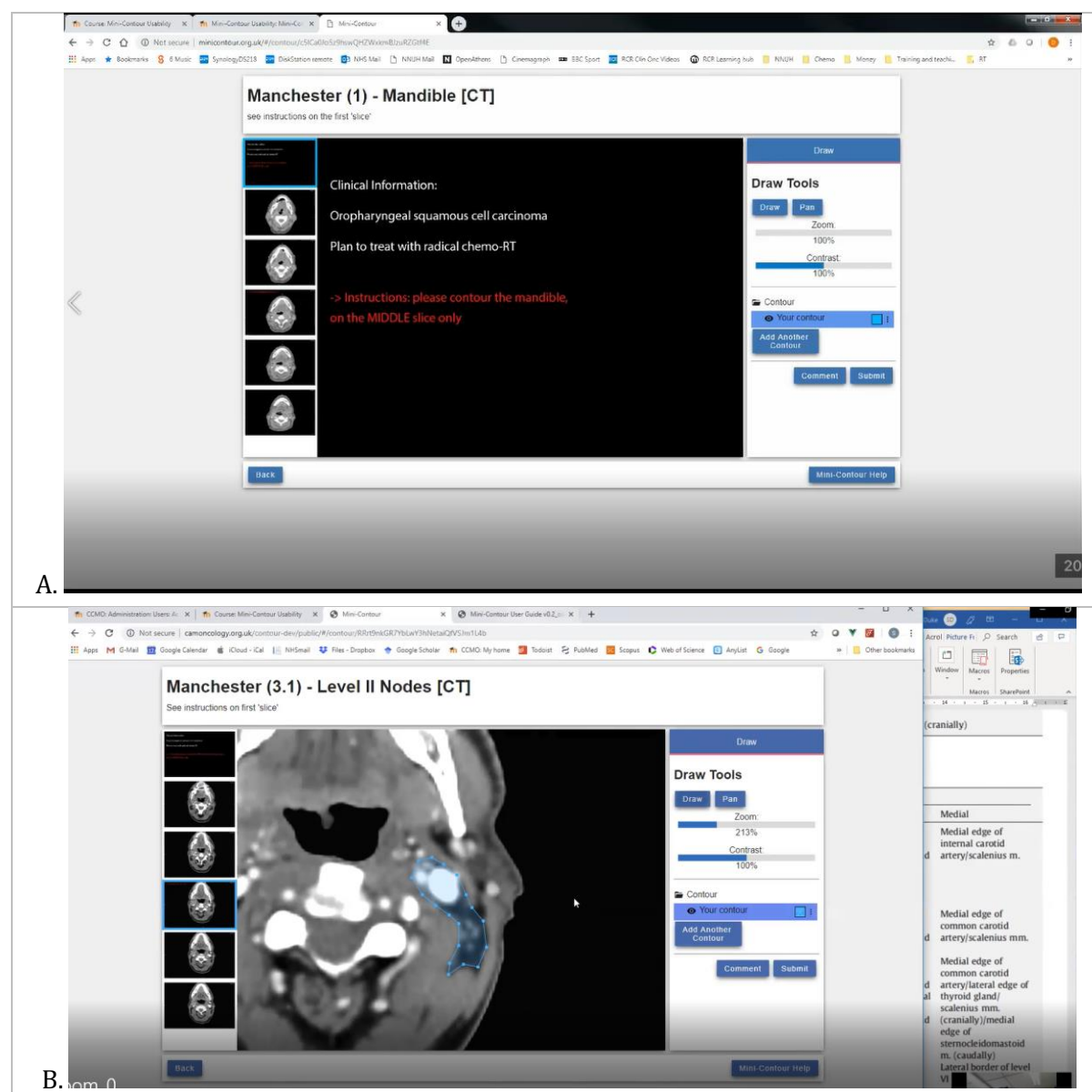


Figure 8-7 - Examples of reduced browser space for Mini-Contour. A: Excess white space caused by browser zoom settings. B: Manually reduced window size to allow space for contouring guidelines

Users were able to quickly understand the zoom, pan and contrast tools: usability issues generally related to using key and/or mouse combinations from their usual clinical software, and initial adjustments were made quickly. One user pointed out the discrepancy between the contrast functionality (which had one intensity threshold) and 'windowing' as used in medical imaging, which has two intensity thresholds (see Figure 8-8 below), for example ST_04: *".. So I can't actually window, I can just adjust the contrast. It's a mandible ... can't get a bony window but let's just go with what we've got"*. Also mentioned were the lack of 3D viewing to help learn cranio-caudal borders, and the lack of image fusion interpretation.

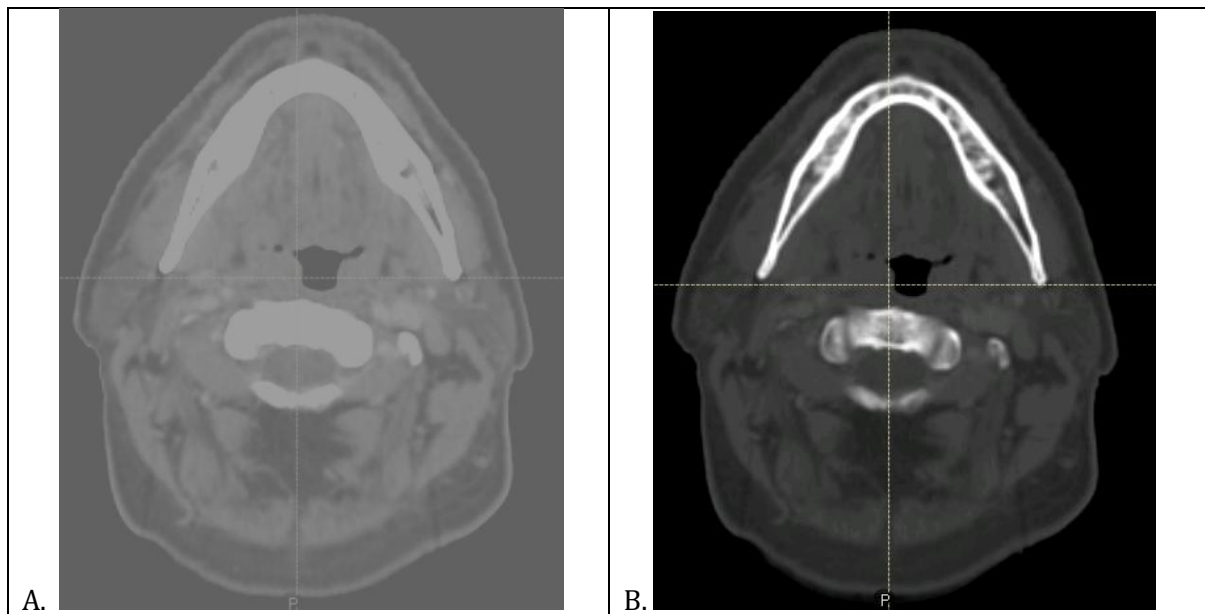


Figure 8-8 - Illustration of the difference between contrast and windowing. A: Mini-Contour screenshot. B: Screenshot of same image on 'bone windows' in radiotherapy treatment planning system

Users varied in their approach to the user guide, with 3/5 reading it thoroughly (>2 minutes); 2/5 briefly opened the user guide, scanned it (for <30 seconds) and kept it open for reference. These two users were both able to solve issues that arose with reference to the guide but missed the 'toggle contour visibility' and 'adding contour points' functions.

Exercise Content

There were a wide variety of issues relating to exercise content (Appendix Table A.8-1). The proportion of issues relating to exercise content increased as the exercise difficulty increased from tutorial (exercise 1), to organ at risk (exercise 2), to elective lymph node CTVs (exercises 3-5).

The one 'severe' usability issue in the study was a result of a junior trainee lacking adequate contouring guidance for the exercise and taking a long time to find an online radiotherapy contouring atlas.

In two instances users drew lateral structures (lymph node CTVs) on the wrong side of the neck despite laterality being specified in the exercise instructions.

Three out of five users identified the deliberate mistake in exercise content (incorrect feedback for final exercise) but only one of these submitted a comment regarding this.

Drawing and editing

Drawing and editing were separate activities in the protocol-specified analysis, but for 3/5 users this separation was artificial as they edited their contours as they drew them. Drawing and editing together caused more 'moderately severe' issues than any other category.

Common issues were inadvertent contour points (n=6), and difficulty adding extra contour points (n=3). One participant mentioned several times that an 'undo' button would be very helpful.

Reviewing feedback

Reviewing feedback caused the highest number of usability issues, although most were mild. Initially users had difficulty relating the colours of the learning zone text boxes, which were coloured according to whether the user had got them right (green) or wrong (red), and the learning zone areas on the image, which were coloured according to whether they were an include (green) or exclude (red) zone:

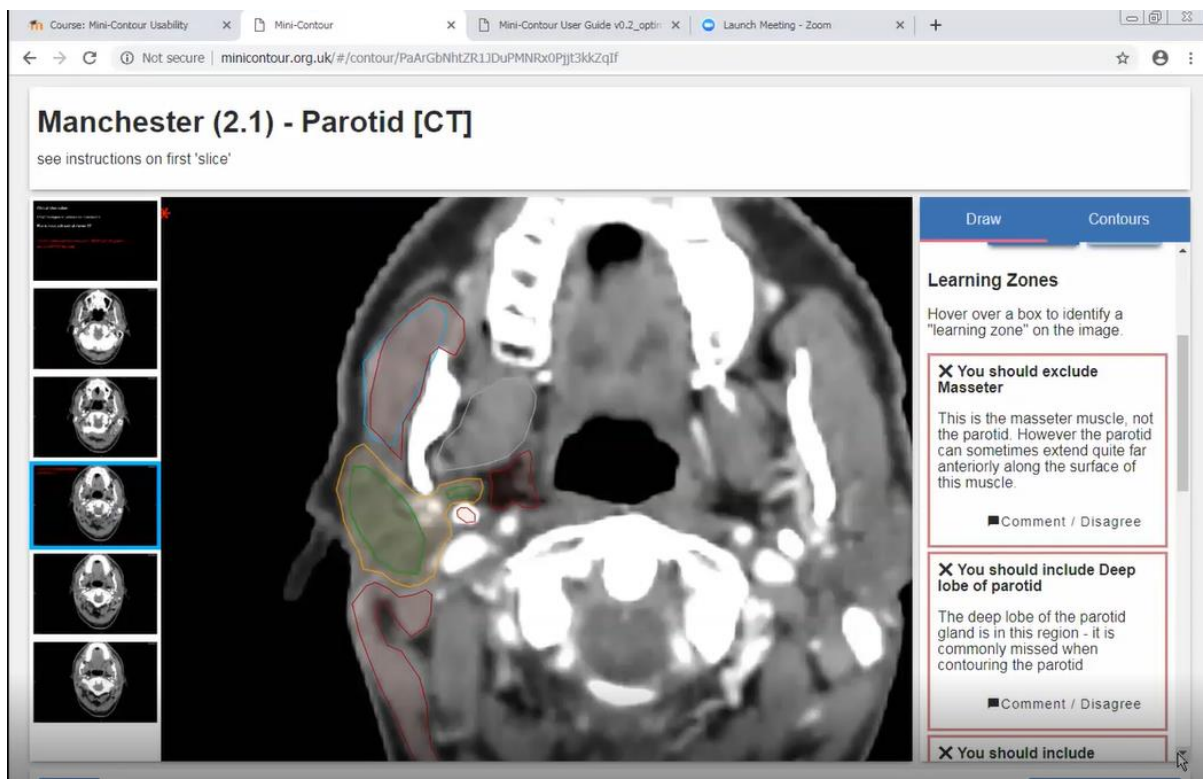


Figure 8-9 - A user's difficulty relating the learning zone written feedback to the displayed areas, due to design / colour choices

JT_01: "I didn't find it 100% intuitive with the right and wrong boxes in terms of matching up ... I think if the colours were matched I would have found that a bit easier, so purple corresponds to this purple blob rather than green and red .. And you had to read it and then look across ...".

Users (especially the junior trainees) sometimes had difficulty interpreting the feedback when a key anatomical landmark (such as the sternocleidomastoid muscle or carotid artery) was not specifically displayed as a learning zone.

Inadvertent scrolling off the learning zone feedback (n=3) and difficulty identifying the gold standard (n=2) were also seen.

8.3.3 Post-task survey & associated themes

This section reports the results of the post-task UTAUT survey (Appendix A.8.5) and the accompanying discussion around its themes. Where relevant, comments made by participants during the exercises or semi-structured interview that pertain to the same themes are included here.

Effect on role

◆ = consultant; ■ = senior trainees; ● = junior trainees

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
I would find the tool useful in my job				■ ◆	● ● ■
Using the tool enables me to accomplish tasks more quickly			■	● ■ ◆	●
Using the tool increases my productivity		■	■ ◆	● ●	
If I use the tool, I will increase my chances of progressing in my career			■ ■ ◆	●	●

Junior trainees predicted a potential time saving, for example JT_01: “it would save the consultants loads and loads of time .. You know when they are first teaching you stuff - especially with head and neck”. All participants noted increased skill was a potential benefit, for example ST_02: “The tool enables me to accomplish tasks more quickly’ .. no, I wouldn't say it was quicker, perhaps just more accurate”.

Most participants envisaged using Mini-Contour early in the course of learning a new tumour site - C_03: “As a consultant [...] if I had to take on a new tumour site .. or if I hadn't done it for a long time I would find this useful reminder .. because your colleagues are so busy”; ST_04: “[it] would be very useful for new SpRs”; ST_02: “I don't think you'd have it open at the same time as contouring a patient, but I think it would be useful to do before you started ...”.

Participants did not see career progression as the end result of Mini-Contour training, rather they focussed on improving skill or quality of care: ST_04: “[it] depends what you think the metric of career progression is .. I think it would improve me as a clinician .. I'm not necessarily sure it would increase my chances of progressing in my career .. I'm going to be neutral on that ..” or appearing prepared - JT_05: “ .. if I'd done this, then do my first nodal outlining - it looks like I know what I'm doing ..”.

Learning curve & engagement

◆ = consultant; ■ = senior trainees; ● = junior trainees

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
My interaction with the tool is clear and understandable				■ ◆	● ● ■
It would be easy for me to become skilful at using the tool			●	◆	● ■ ■
I would find the tool easy to use				■	● ● ■ ◆
Learning to operate the tool is easy for me				● ■	● ■ ◆
Using the tool is a good idea				■	● ● ■ ◆
The tool makes learning more interesting				■ ■	● ● ◆
Working with the tool is fun		■ ■		● ● ◆	
I like working with the tool				■ ■ ◆	● ●

Mostly users reported finding Mini-Contour easy to learn to use: C_03: “Yeah I think it's mostly really good, it's a quick learning curve .. you can quickly learn to navigate it & learn where the tools are ... it's intuitive ..”; JT_01: “It didn't take long to figure out how it was going to be helpful”.

At some point during the session, all participants compared and/or contrasted Mini-Contour with the contouring software that they used in their routine clinical practice, for example: JT_01: “I think it would take a little bit of time, just because some of my reflexes are wrong .. Because it doesn't mirror exactly how you contour on Raystation [software]”; JT_05: “Intuitively I would have thought I can do it like I can do it in Prosoma [software] by shift clicking or alt clicking .. But none of that works ..”.

Two participants commented on the role of 'fun' in professional training: ST_02: "It's not fun ... it's still work [laughs]"; ST_04: " 'Fun' - that's an interesting word .. I'm going to be neutral because I'm not sure that fun is really what I want in a tool like this. Useful rather than fun".

External influences

◆ = consultant; ■ = senior trainees; ● = junior trainees

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
People who influence my behaviour think that I should use the tool			● ■ ■		●
People who are important to me think that I should use the tool			● ■ ■ ◆		●
The senior management of my organisation have been helpful in the use of the tool		■	● ■		●
In general, my organisation has supported the use of the tool			● ■ ■		●

'External influences' were felt by most participants to be muted or non-existent as they were unaware, for example - ST_04: "I don't know anyone else who actually knows about the tool .. I've never talked to anyone about it apart from you"; C_03: "Well they don't know about it yet but I expect that if it existed they would [want people to use the tool]". One participant (JT_05) responded hypothetically in this section.

Pre-requisites & compatibility

◆ = consultant; ■ = senior trainees; ● = junior trainees

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
I have the resources necessary to use the tool			●	■	● ■ ◆
I have the knowledge necessary to use the tool				● ■	● ■ ◆
The tool is <i>not</i> compatible with other systems I use	● ◆	● ■	■		
A specific person (or group) is available for assistance with difficulties with the tool			● ■	◆	● ■

User who completed the test on their own computers were concerned with compatibility with NHS systems , for example: JT_05: “... needs to run on outdated browsers which are in common use in the NHS”, but this was not an issue for the two who actually used NHS computers. Two participants noted that Mini-Contour doesn’t interact with current radiotherapy systems (e.g. clinical contouring software) i.e. it stands apart as a training tool.

Participants found the UTAUT questions on task completion “with assistance confusing, partly due to the observed nature of the test, and partly as they had completed the exercises autonomously already.

Anxiety, safety & self-concept

◆ = consultant; ■ = senior trainees; ● = junior trainees

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
I feel apprehensive about using the tool	● ■	■ ◆		●	
It scares me to think that I could lose a lot of information by hitting the wrong key	● ■	● ◆	■		
I hesitate to use the tool for fear of making mistakes I cannot correct	● ● ■ ◆	■			
The tool is somewhat intimidating to me	● ● ■ ■ ◆				

Most participants didn’t report feeling anxious using the tool. During the exercises, there were some spontaneous expressions of anxiety and/or threat to self-esteem, for example - JT_05: “Right now I feel stressed .. because it's like an exam .. that I'm failing”. In these situations it was difficult to disentangle the anxiety of assessment by Mini-Contour from that of being directly observed - JT_01: “ [reading] ‘I feel apprehensive about using the tool’ .. a little bit .. but probably not if I wasn't being filmed doing it”.

Trainees touched on psychological safety of low-stakes simulation during their exercises: JT_01: “It's a safe space to make errors, really”; JT_05: “I feel like it's actually quite nice .. Because I can give it a go and then actually just find out how far [away] this was .. probably quite a bit!”. P02: “Well I'm not exactly sure, but that's the point of this tool .. so that's fine - I'll just have a go ..”.

Plans for use

◆ = consultant; ■ = senior trainees; ● = junior trainees

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
I intend to use the tool in the next 12 months		◆		● ■ ■	●
I predict I would use the tool in the next 6 months			◆	● ■ ■	●
I plan to use the tool in the next 6 months		◆		● ■ ■	●

Plans for use were difficult for participants to predict as the tool is not widely available. The consultant reported they would recommend it for others but not plan to use it themselves.

8.3.4 Other themes & semi-structured interview

Learning zones

The recordings show multiple instances of learning zones providing corrective feedback on contouring decisions, for example - C_03: “Oh!! I was right the first time .. I shouldn't have changed it! .. That's still part of sternocleidomastoid”; ST_02: “Ok ... [reads feedback re: masseter] .. Ok, so I've gone slightly too far anteriorly ..”. ST_04: “[Hovering over digastric learning zone] You should exclude digastric, and I did not exclude .. I went right up to the submandibular .. Bad me .. But that was an anatomy issue in identifying digastric”. One user reflected (JT_05): “This is a lot easier than doing it on actual patient cases, because ..(.).. you actually know 'this is wrong because you have included this vessel'. You do it in a patient who looks different - you might not actually know why it was wrong”.

However feedback was sometimes confusing if it was incomplete or the overall impression was overly negative. When a key anatomic landmark was not included as a learning zone users sometimes made incorrect inferences about their errors, for example - C_03: “Ok, so it says to contour the vessel but I thought the big one was the vessel and I included that. It does say medical border and I included that, so I'm not sure .. [user reviews 'vessels' learning zones but internal carotid artery and internal jugular vein are grouped as 'vessels']”. At times however, users did manage to make correct inferences about the difference between their contour and the faculty contour in the absence of learning zone feedback: (ST_02) “Sorry I just want to figure out which one was the common carotid .. Ah ok, I think that's bone isn't it .. I think this one is the carotid artery”.

Complaints about learning zone stringency - i.e. users being marked incorrect for small areas of unsatisfactory overlap - were relatively common, but users were often able to make their own inferences if they felt that they had been harshly judged, for example - JT_01: “You should exclude scalene .. Which .. I did, no? Ah, yeah, harsh harsh, I've only got like 1 mm there. I think mine's better than the answer [jokingly]”. Another example is shown in Figure 8-10:

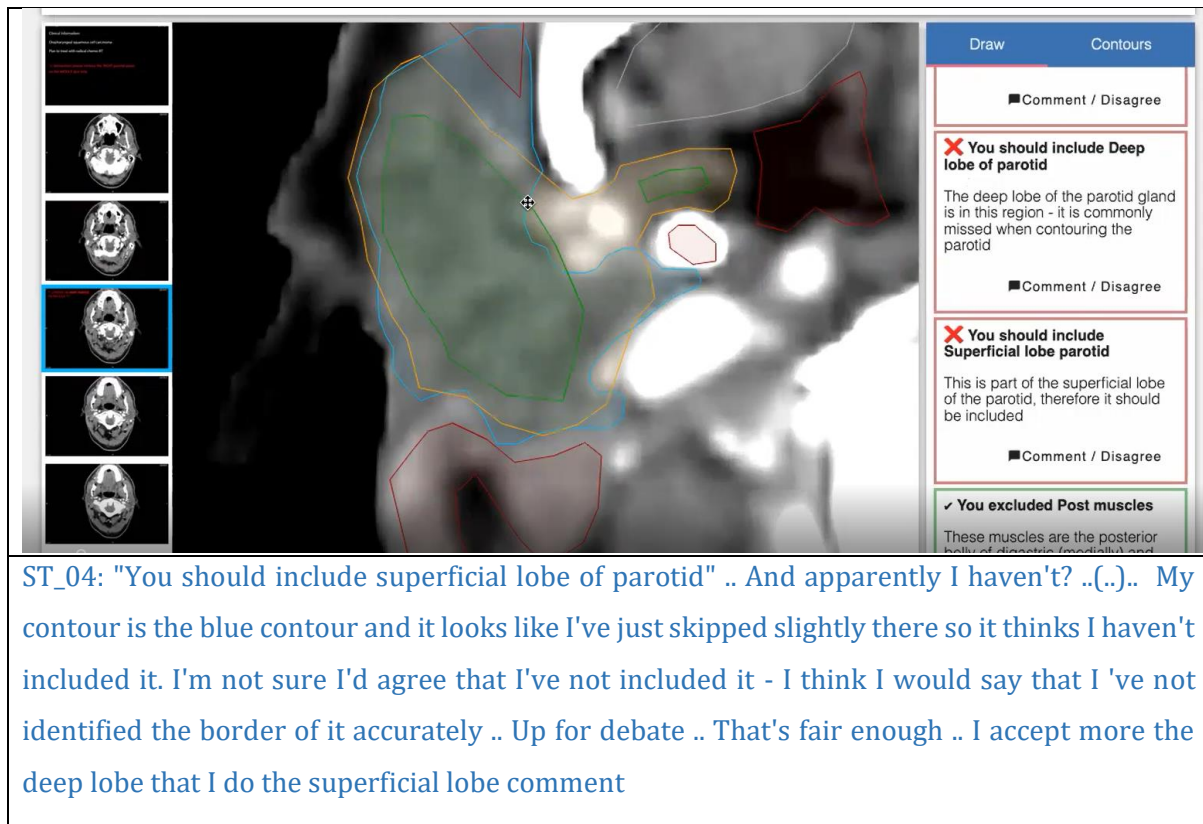


Figure 8-10 - A user reviewing learning zones and debating the stringency of automated assessment: screen shot and associated transcript

Sometimes users' reaction to their performance was negative (for example: “[Exclaims] Oh! What have we done - we've done something wrong.. Shoot! .. (..) .. ok, this is upsetting - everything has been major errors here ...”) even when their overall performance was very good, due to the presence of an ‘incorrect’ exclude learning zone displaying first. More time was spent attending to negative learning zone feedback than positive - attention to which sometimes seemed cursory.

At some point all participants reflected on the impact of the learning zone feedback on their psychological state, for example C_03: “But mostly right so I'm happy because I've got more green squares than I did before .. And no red squares as well .. Happy days!”, JT_05: “[smiles - ‘successfully’ excluded some organs at risk from the lymph node volume] so I find it motivational that although it's completely wrong at least I've done something right” (Figure 8-11):

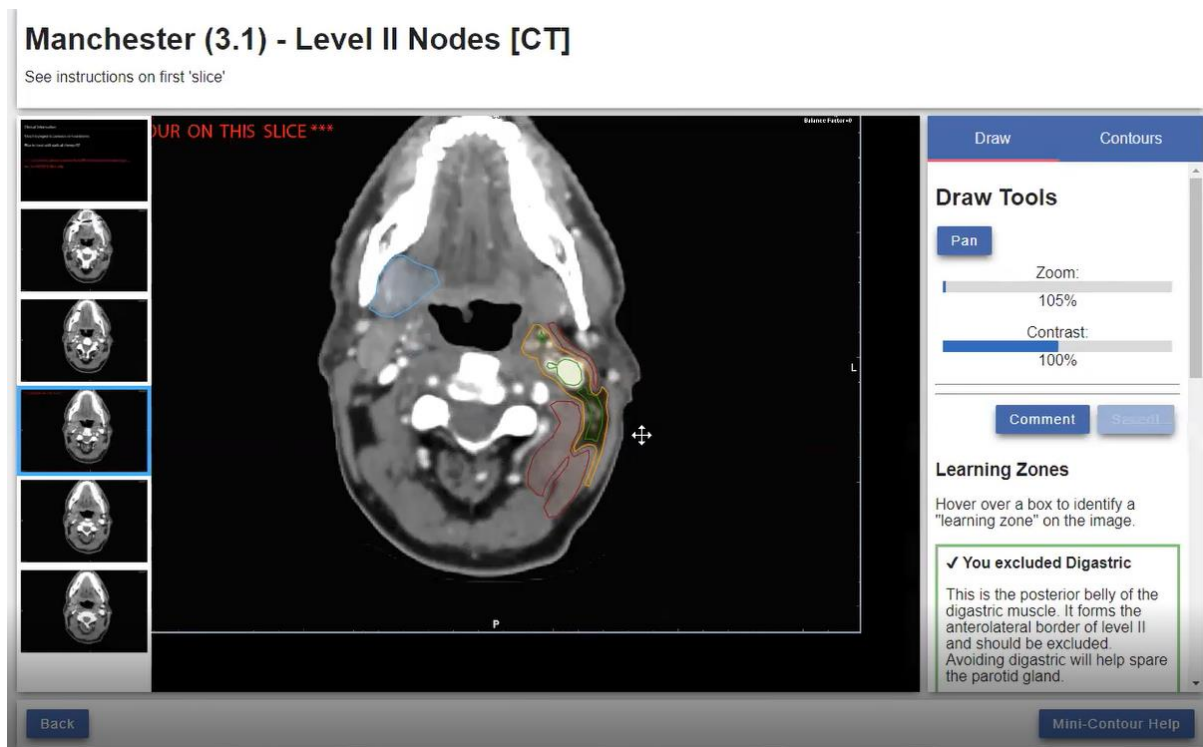


Figure 8-11 - An example of a user evaluating learning zone feedback: "I find it motivational that although it's completely wrong at least I've done something right". User contour in blue and faculty contour in gold.

Contouring atlas & other resources

In total, 2/5 participants spontaneously searched for (and found) a relevant contouring atlas. The three other participants all made references to contouring atlases. One participant specifically contrasted the learning with and without help from an atlas - JT_05:

"So obviously you could see a contouring guide right next to it, but it just doesn't go in .. I've looked at the axillary contouring guide so many times for breast .. (.). Because I make the mistake I look at what it means. [Using an atlas] I would just copy the structure and be right - I wouldn't bother reading all the comments. Now I've actually thought a little bit about - 'what is the internal carotid?'. 'Why is the internal carotid not bright white?' So I feel like I engage more with it if I make mistakes .. go back and look at why I've made the mistakes, even though I made some quite significant mistakes you could say I've actually engaged with this a lot more ..".

8.3.5 Emergent themes

As part of the inductive process of coding, themes regarding users' cognitive processes emerged. Two of the most prominent were reasoning processes and self-regulation. Data are presented below and explored further in the discussion.

Clinical reasoning processes

Users spent a significant amount of time relating the contouring guidance to the anatomy of the individual patient, often swapping views between the contouring guidelines and the case. They often used landmarks such as bones, muscle or blood vessels to help them make sense of the individual patient's topography, for example - JT_01: "Ok, so it's all about sternocleidomastoid and common carotid artery .. So if I find those two things then I'll know where I'm starting ...".

Scrolling through the slices above was seen as a method of checking that they had identified a given structure successfully, which was sometimes limited by the number of slices: ST_02: "[scanning up and down the images] I guess I'm struggling a bit with which one is the vessel - I normally go a bit further down than this. So I normally follow it all the way down to the main branch. I guess especially with head and neck it is helpful to see a bit higher up and further down".

Mentions of patterns of tumour spread guiding contouring were relatively rare (3 comments in total).

No participants vocalised thoughts about the dosimetric consequences of their contouring despite at least 3 learning zones referring to the effects of contouring or excluding specific region on dosimetry.

Self-regulation / Meta-cognition

Several self-regulated processes (Sandars and Cleary, 2011, de Bruijn-Smolders et al., 2014) were evident in participants' thoughts including mental rehearsal, monitoring, attribution, and adaptive interference.

Self-regulated processes before submission included **mental rehearsal** (trainees tracing their contours before starting), and **monitoring** - ST_04: [has closed the contour and is looking at it] "Alright ... so my areas of concern here are .. (..) .. I think that's as deep as it [the parotid gland] goes but let's have a look on some other slices to reassure ourselves"; (C_03) "So ... I'm going to remove these .. This is where I'm least confident" [deletes contour points and expands contour forward towards submandibular gland].

Attribution - causal retrospective inferences about performance - was also seen in response to feedback, with or without the presence of a relevant learning zone. An example where a learning zone was not provided is - ST_02: "So I think maybe I've included the submandibular gland [reads again] .. So I think I've gone too far forward there ... right so I think I should be ..". Examples of self-evaluation after learning zone feedback include: ST_04: "Ok - so lesson there was deep lobe of parotid .. the lesson here is get digastric right and refer to an atlas when you can't see it" & P05:

“[reads] ‘you included the common carotid’ .. Ok .. I should have noticed that this is venous contrast and that means the veins are brighter ..”.

Adaptive interference, i.e. modifying one’s learning strategy to suit the situation, was displayed by the junior trainees. One trainee (JT_05) repeated each exercise after a preliminary attempt; similarly the other junior trainee’s attitude (JT_01) was to “give it a go, and then I’ll learn from my mistake”. That trainee did not repeat the exercises formally but on two occasions ‘traced’ a new contour on a different image.

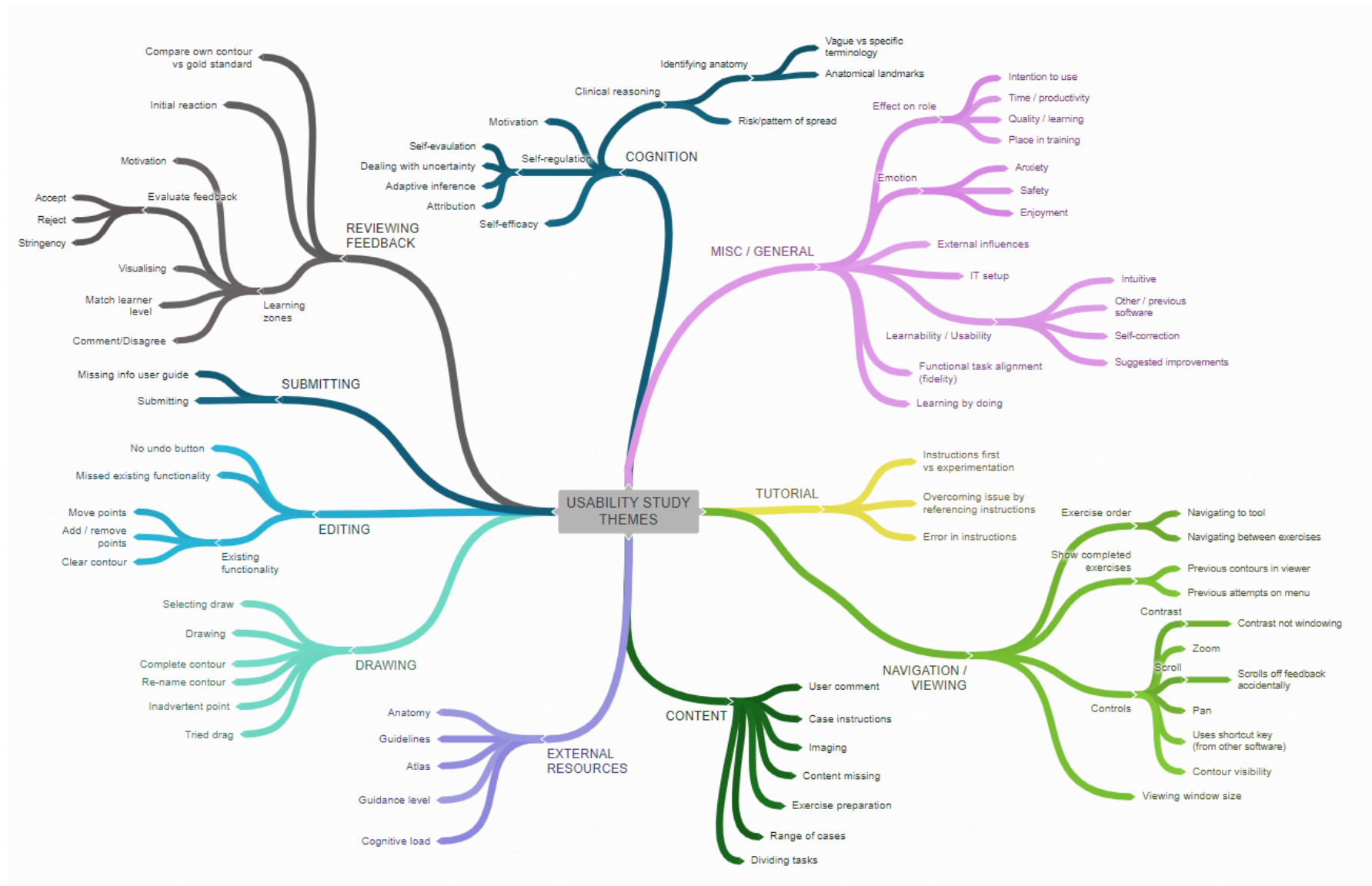


Figure 8-12 - Thematic map of users' comments in Mini-Contour usability study

8.4 Discussion

Mini-Contour has been shown to operate with users' workplace IT setup (in the NHS) as well as with their personal IT equipment, on laptop and desktop computers, and using a mouse or laptop trackpad for contouring, albeit for a very small numbers of users.

Many usability issues stemmed from users' instincts or habits from using other software which need to be 're-programmed' - the declining rate of usability issues with each exercise shows that this was mostly achievable.

Insight has also been gained insight into users' reaction to feedback, clinical reasoning & self-regulation processes.

User guide

As the users who skimmed the 'user guide' missed relevant functionality, it may be helpful to setup an interactive tutorial case where users work through an exercise completing specific functionality step by step - this is common in the software programming world ^x.

Drawing and editing

The moderate and relatively persistent usability issues seen for drawing and editing necessitate a significant update in functionality. A 'rollerball' or 'brush' with both drawing and editing capabilities would require more complex programming, but may be worth the investment if it is a significant part of learners' usual workflow as this would improve functional task alignment (see Chapter 3 section 3.4.2). 'Drag to draw' contouring functionality is less of a priority as it is difficult to perform on a laptop trackpad and users were able adapt quickly to draw clicking to make points.

Understanding users' previous contouring software experiences and habits to optimise is clearly important in guiding development. Of note, 4/5 participants had trained in Cambridge University Hospitals where the radiotherapy treatment planning software (Prosoma®) default drawing functionality was 'click to draw'. Further sampling of clinicians with experience of different radiotherapy software is therefore required.

^x for an example, sign up to the <http://codecademy.com> beginners' HTML course

Exercise content

The increasing proportion of issues relating to content as exercise difficulty increased points to a need to provide guidance appropriate to the level of the learner. The lack of an atlas or formal instruction for the inexperienced trainees caused them to struggle. As noted in Chapter 3, worked examples or pre-learning can both reduce cognitive load and improve novices' learning. Flicking frequently in between the guidance (i.e. comparing two sources of information) as seen in this study may be increasing 'extraneous load', although it could be seen as intrinsic *if* it enhances learning. As one participant suggested, learning without an atlas may promote learner engagement, or in the language of cognitive psychology "active processing with schemata formation in long-term memory". This should be balanced against the proven ability of radiotherapy atlases to improve contouring quality and reduced variation (see Chapter 2).

Advanced learners benefit from less support i.e. a reduction in 'scaffolding' (see Chapter 3, Section 3.5). Further work to explore the effects of changing levels guidance to suit the learner, and the presence or absence of an atlas, is warranted. Functionality to facilitate individualisation to the level of the learner should be an aim of future Mini-Contour development, although this is not straightforward technologically. This approach signifies a shift in focus from the individual learning exercises to the overall 'learning path' towards mastery, and also requires consideration of the overall curriculum.

Learning zones

The initial high frequency of usability issues in users reviewing the feedback is unsurprising given that the 'learning zone' concept is unfamiliar. Immediate development should focus on more clearly linking the learning zone feedback text with its corresponding region - simply changing the box colours, so that the borders match the learning zone contour colours, is likely to help.

A conflict between the use of learning zones for assessment and their use for feedback was apparent, especially for exclusion zones. Exclusion zones more commonly took the shape of an anatomical structure (e.g. a muscle), which were then cut back a distance from the gold standard contour to allow a margin of error. Users generally seemed not to realise that allowances had already been made. In some cases these allowances turned out to be insufficient to account for acceptable variation.

It was notable users generally were able to draw their own conclusions about whether their own variation was acceptable even if their contour was assessed as breaching an exclusion zone, however the irritation produced by apparently over-stringent assessment may detract from the

credibility of Mini-Contour feedback. The stringency, and sensitivity to small areas of overlap, of learning zones will be explored further in Chapter 9.

A potential solution to this conflict is to separate ‘assessment’ zones (which would be hidden from the user) from ‘feedback’ zones (which would be displayed). The assessment zones could be either automatically trimmed back from the feedback zones, or a separate assessment zone (in the most critical area) created manually (Figure 8-13).

The dismayed user reviewing what appears to be negative feedback for generally an acceptable contour (due to an “incorrect” overly stringent exclusion zone appearing first) is counterproductive - especially if we are wanting to promote target coverage first and foremost over sparing of organs at risk. For exclusion zones Mini-Contour assessment does also not take into account the clinical significance of any variation although it may be explained in the written feedback. Including a small area of muscle (of no clinical consequence) may appear on first inspection to be as grave an error as unnecessarily irradiating an entire salivary gland (which could cause a permanently dry mouth). Giving any positive feedback to a user who has contoured a salivary gland on the wrong side of the neck instead of the lymph node CTV (see Figure 8-11), even if well-received, seems somewhat peculiar.

Inclusion zones should therefore be prioritised in the feedback display to signify the importance of tumour coverage. Many structures that need to be excluded could be created as unassessed ‘comment’ structures (or weighted for their clinical significance), leaving exclusion zones for clinically significant organs at risk, for example the spinal cord or salivary glands (in the head and neck) or bladder or bowel (in the pelvis).

The low number of formal comments regarding learning zones despite users vocalising disagreements with their stringency and/or placement relatively frequently (including noting the deliberate error) suggest that formal comments logged in the tool should be viewed as the ‘tip of the iceberg’ of learner disagreement.

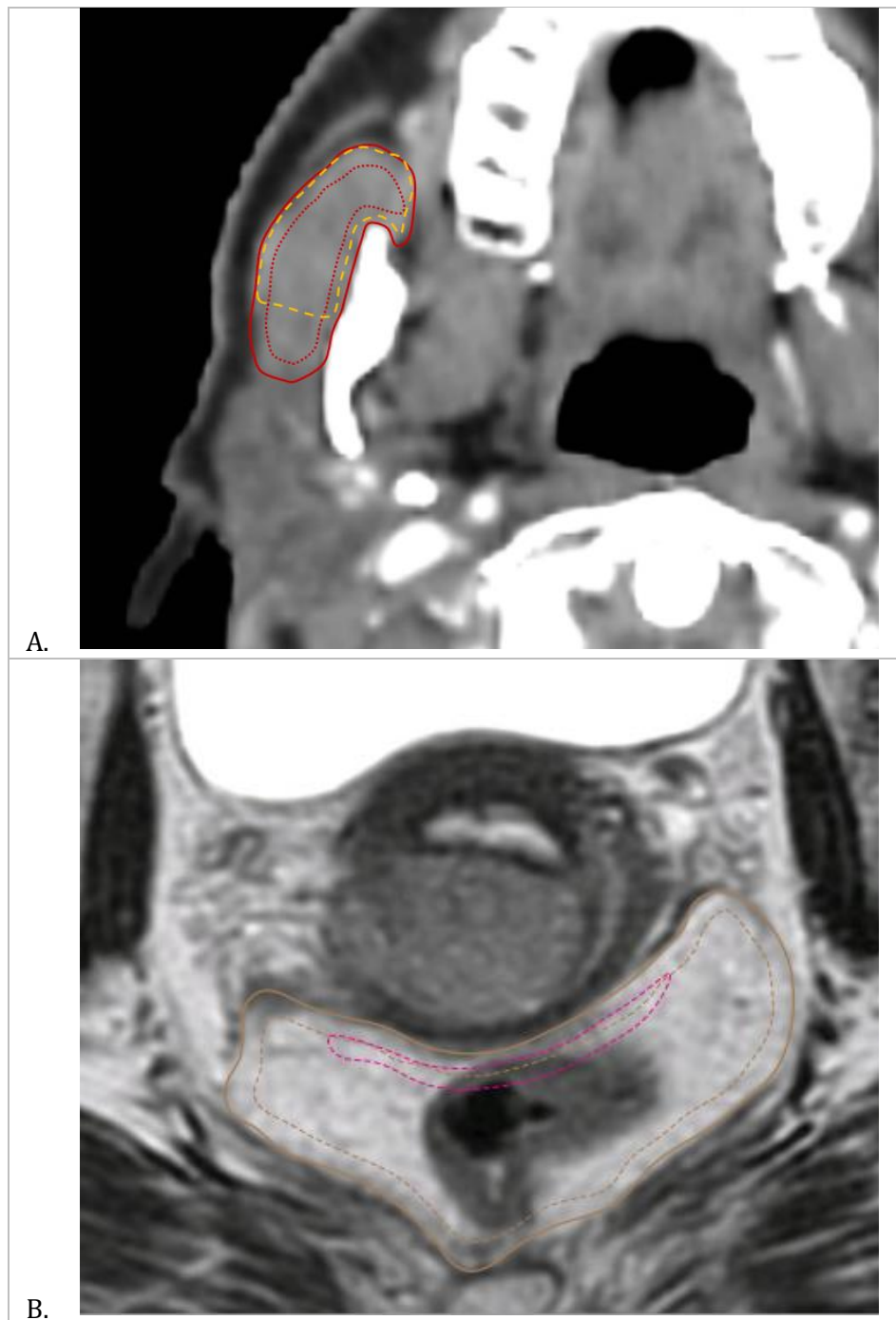


Figure 8-13 - Learning zones: separating feedback from assessment. **A (Head & neck):** Masseter muscle is the red line - this would be displayed to the user. Dotted red line is the automatically cropped 'assessment zone'; dashed orange line is manually drawn alternative. **B (Cervix):** Anatomical mesorectum is brown line - this would be displayed to the user. Dotted brown line is the automatically cropped 'assessment zone'. Dotted magenta line is an example of a manual 'assessment zone'.

Place in training

Participants generally envisaged Mini-Contour would be deployed early in the process of learning a new speciality, although the exercises didn't include difficult cases where there is disagreement amongst practicing clinicians.

The UTAUT questionnaire highlights that motivation via institutional and peer networks is an important factor predicting future use where Mini-Contour is currently lacking. Simulation training programmes in surgery, though available at minimal cost, have struggled with engagement: available time, competing commitments, and interest from trainers being key factors (Gostlow et al., 2017, Blackhall et al., 2019). Dedicated time for simulation training (in parallel with gaining clinical experience) and increased extrinsic learner motivation could be facilitated by buy-in from professional bodies and local training directors - engaging these stakeholders would be key to a learning programme that is well utilised.

Clinical reasoning

Analysis of reasoning process in this group revealed users mostly grappling with anatomy and guideline interpretation with very few references to risk or pattern of spread, and none whatsoever of dosimetric effects. Errors were triggered by anatomical misconceptions which underlines the importance of anatomy teaching recognised by the Royal College of Radiologists (Royal College of Radiologists, 2020b) as well as much of the contouring literature. It is possible that novices find it difficult to move beyond anatomical landmarks when they are uncertain, whereas experienced clinicians are more aware of dosimetric and clinical implications - this hypothesis would need to be explored as part of a dedicated study. Clinicians' (especially junior trainees') mixed ability to correct or explain their errors hints at the potential for automated qualitative feedback to improve self-evaluation.

This study has shown the potential of low-fidelity simulation to capture clinical reasoning processes, but is limited in its inferences by the small sample size and lack of expert clinicians. Virtual patient simulations are a well recognised tool for exploring the development of clinical reasoning processes (Berman et al., 2016). Further study of reasoning processes in simulated and naturalistic settings, in both novice and experienced clinicians, may provide further insight into reasoning processes and how they facilitate expert performance or lead to errors. The time taken to transcribe and analyse clinicians' vocalised thoughts and actions in this small study was substantial - such a study may be facilitated by computer-assisted qualitative data analysis software and should be separated from usability testing which can still be productive without full transcription and such in-depth analysis (Barnum, 2011).

Self-regulation

Self-regulation is defined by Zimmerman as: *“processes whereby learners personally activate and sustain cognitions, affects, and behaviours that are systematically oriented toward the attainment of personal goals.”* (Zimmerman and Shunk, 2011, p.1). Research in this field has shown that self-regulatory process can partially account for differences in achievement amongst learners in higher education (Richardson et al., 2012). Improving self-regulation is also effective in improving performance for learners with a range of baseline ability, but is particularly applied in remediation (Sandars and Cleary, 2011).

The presence of self-regulatory processes such as mental rehearsal, monitoring, process attributions and adaptive inference is perhaps unsurprising in this self-selected group of high-achieving professionals. One cannot draw any firm conclusions from the data presented about the relationship between self-regulation and learning contouring, or of the impact of simulation and/or feedback on self-regulation. However these initial data highlight the importance of, at the very least, being aware that rapid practice and automated feedback may influence self-regulation, whether positively or negatively.

Limitations

The validity of the study is limited by the small sample, homogeneous training background, and limited geographic territory of the participants. Although some broader usability data can be gathered from larger cohorts of learners (see Chapter 9), it is important to complement that with more detailed qualitative insights in the next iteration of this study, especially from site-specialist radiation oncologists, international clinicians, and those with different experience of contouring software.

The usability sample size was partly limited by the significant time taken to fully transcribe and analyse the video recording. Full transcription could be dropped for future sampling with a narrowing of the study focus on usability issues alone (rather than more broad themes such as meta-cognition and clinical reasoning). Unsupervised testing with or without user commentary can be provided by 3rd party software (for an example see <https://userbrain.net/>).

I knew four out of five participants professionally; it does not take great reflexivity to realise that this may have influenced their actions and expressed views. It is therefore important to triangulate these qualitative findings with other cohorts, ideally with an independent interviewer, although this may be impractical without substantial extra funding or collaboration. Similarly, coding and content analysis was performed by a single researcher (myself). Coder bias

could be significantly reduced by multiple coders but this would also result in an increase in the resources required.

Some usability data, such as the time taken per exercise, was not possible to collect due to the interview format and think aloud methodology. This information can be gathered from more naturalistic educational settings and will be presented in the results of the pilot studies in Chapter 9.

Heuristic evaluation (i.e. review by an independent usability expert) was not conducted in this iteration. Heuristic evaluation is not commonly used in usability assessments, probably because of the requirement for domain experts to perform it (Yen and Bakken, 2012), but would likely supplement future evaluations if the relevant expertise can be secured. Focus groups enable researchers to obtain rich insights from a broader audience than individual interviews allow (Cohen et al., 2017), and would also be appropriate for future development cycles.

Implications for further development

Implications and plans for future development are discussed in Chapters 9 & 10 where the results of wider user testing are collated with the findings from this study.

8.5 Conclusion

Despite limited sampling, this study has discovered multiple usability issues, many of which can feasibly be addressed in the next development cycle. Users adapted to the software quickly and reacted positively to the novel elements of rapid practice and learning zone feedback. They envisaged this would be most useful when learning a new tumour site specialism. Clinicians' ability to make their own inferences about contouring errors was mixed, especially for junior trainees in the absence of specific anatomic landmarks.

The study also highlighted the importance of broadening the design perspective beyond the simulation software and learning exercises, to include: a contouring curriculum and learning path to mastery (ideally adapted to learner's ability), the interplay between simulation and users' reasoning and self-regulatory processes, and engagement of professional networks vital to the uptake of any learning programme.

The priorities for future development will be outlined after these data are combined with usability data from larger cohorts (Chapter 9). Further iterative cycles of development and evaluation are required, especially with site-specialist clinicians; these could be conducted

separately from (and likely more efficiently than) in-depth studies of clinical reasoning and self-regulation processes.

9 Teaching contouring using Mini-Contour: three pilot studies

9.1 Introduction

Although the in-depth usability data reported in Chapter 8 are valuable to guide future development of the Mini-Contour simulation, no educational design research programme would be complete without an evaluation of the intervention *of its effects on learning* in its intended learning environment as part of the ‘evaluation & reflection’ phase (McKenney and Reeves, 2020). This type of study is also known as a ‘design experiment’ or ‘formative evaluation’ (Cobb et al., 2003, Nieveen and Folmer, 2013) and provides data to further refine and develop the prototype. Such a study also allows evaluation of usability and acceptability in a broader population than is feasible in a detailed usability study.

This chapter comprises three pilot studies:

- A one-off workshop with UK trainees
- A longitudinal programme with international trainees
- A one-off workshop with EMBRACE group clinicians

All three pilot studies were conducted under the same protocol, so are reported together in this chapter.

9.1.1 Aims and research questions

Aims

The overarching aims of these pilot studies were to explore the feasibility, acceptability, and usefulness of the Mini-Contour simulation for teaching contouring to groups of oncology trainees and accredited clinicians. Study endpoints are shown in Table 9-1:

Table 9-1 - Mini-contour pilot study primary and secondary endpoints

Endpoint	Domain	Sub-domain	Measurement
Primary	<i>Feasibility</i>	Time taken	Time taken to deliver each learning exercise Time taken to create each learning exercise
Secondary	<i>Feasibility</i>	Exercise creation	Faculty agreement with learning zones
		Learning environments	Number of exercises & time taken in each session Types of environments evaluated
		Engagement	Proportion of learners completing learning programme & follow-up exercises
	<i>Acceptability</i>	Satisfaction	Reported satisfaction
		Usability	System usability scale
		Perceived fidelity	Scale of 0-100% similarity to real life
		Free-text comments	Content & thematic analysis
	<i>Usefulness</i>	Perceived usefulness	Reported usefulness & relevance of learning zones Reported enthusiasm for future use
		Confidence	Reported confidence at baseline and after learning exercises
		Performance	Performance at baseline, during learning programme, and at delayed follow-up - measure by conformity index and learning zone success rate

Research questions

- How does the perceived usability and fidelity of Mini-Contour vary across groups of different experience and expertise? How do clinicians think this can be improved?
- What are clinicians' perceptions of the automated feedback?
- Are the errors seen in high-fidelity simulation replicated in low-fidelity simulation (Mini-Contour) across different cases and cohorts?
- How does performance (assessed by conformity index and learning zones) relate to confidence and clinical experience?
- How do confidence and performance change over time?
- Do experienced clinicians perform better on Mini-Contour?

My contribution

This study was the result of a collaboration with multiple groups to facilitate recruitment to the study and data collection. Collaborators are listed in Appendix Table A.9-1. I led the study design,

ethical approval, data collection & analysis, and wrote the results and discussion presented below.

9.2 Methods and materials

All three pilots were conducted with Mini-Contour version 1.0 - the same version tested in Chapter 8.

9.2.1 Mini-Contour exercise creation

Initially, 16 cervix cancer contouring exercises were prepared for the trainee EBRT workshops by a single investigator (SLD); they focussed on contouring of the GTV, local CTV^{xi}, and elective lymph node CTV for EBRT.

Learning zones were created based on difficulties seen in the EMBRACE-II (Chapters 5 & 6) and INTERLACE (Eminowicz and McCormack, 2015) radiotherapy quality assurance exercises, as no data were available regarding contouring difficulties encountered by trainees. A second and then third investigator (GE, RN) then contoured independently and discrepancies were reviewed, with adjustments made to learning zones as necessary (Figure 9-1). Editing cases required manual alteration of the underlying database as Mini-Contour v1.0 lacked dedicated case editing functionality.

The time taken for all these steps was recorded and presented in ‘person minutes’ i.e. two faculty reviewing contours for 20 minutes counted as 40 minutes.

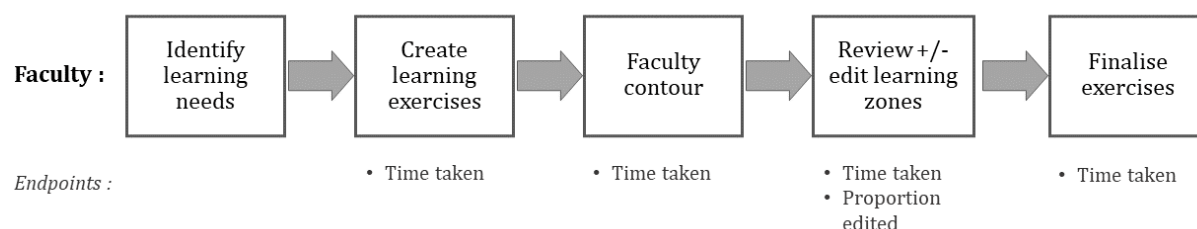


Figure 9-1 - Flowchart of Mini-Contour exercise creation and associated study endpoints.

Each exercise contained several different learning zones. Some learning zone ‘themes’ were re-tested in the trainee workshop(s) and/or follow-up exercises, for example inclusion of the left

^{xi} i.e. relating to the primary tumour. This has different names in different cervix cancer protocols. GYN consensus: “CTV”, EMBRACE-II: “Low Risk CTV-Tumour”, INTERLACE: “CTV1”

lateral para-aortic lymph node region in the elective lymph node CTV, or exclusion of the mesorectum from the local CTV.

Subsequently, additional exercises were created for the EMBRACE workshop (3 brachytherapy and 1 EBRT) and international trainee longitudinal programme (10 EBRT).

9.2.2 Study cohort & procedures - UK trainee workshops

The study was conducted as part of 3 separate regional training days in South London, North London and Manchester. All three training days contained lectures on the anatomy, radiology and contouring of cervix cancer prior to the contouring workshop. In total, 85 trainees were invited to participate in the study of whom 80 were included in the study after giving consent, completing a survey and more than 1 contouring exercise as per the protocol.

At the start of the workshop (Figure 9-2), trainees completed an online questionnaire collecting information about their stage of training, cervix cancer experience, and confidence in contouring relevant radiotherapy target volumes and organs at risk (1 = not at all confident, 5 = very confident - see Appendix A.9.2).

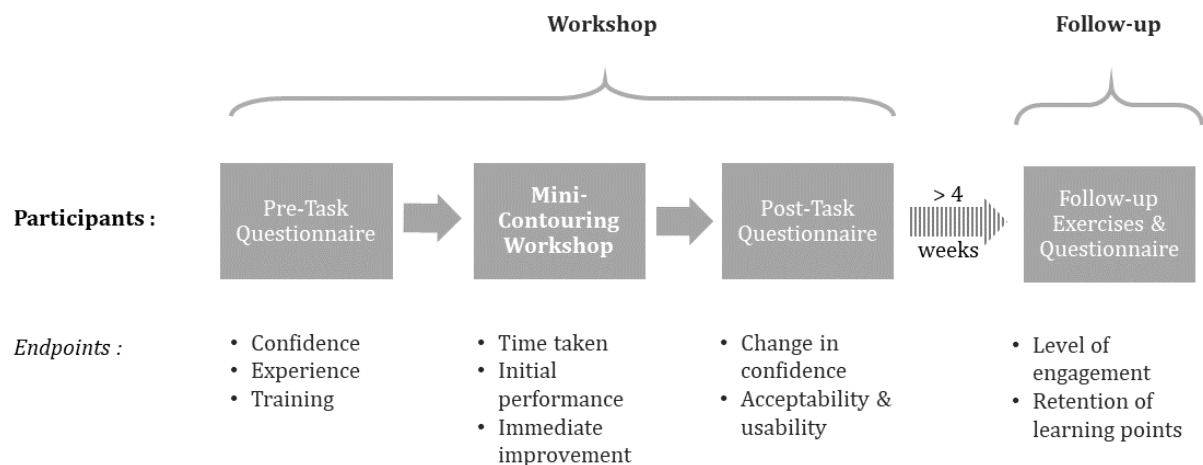


Figure 9-2 - Flowchart of UK trainee workshop study procedures

Trainees worked through a tutorial case with the workshop facilitator (SLD) to become familiar with the simulation interface. They then completed 12 contouring exercises: 6 Elective lymph node CTV exercises (on CT) followed by 6 exercises on anatomy, GTV contouring, and local CTV (all on MRI). After each exercise, the group's results were discussed briefly by the workshop facilitator (SLD and local faculty). The time taken for each exercise (loading, contouring, feedback and reflection) was recorded within the tool. Trainee performance was assessed by Jaccard

conformity index and appropriate inclusion/exclusion of each learning zone in its entirety (i.e. a binary score for each learning zone).

In all three pilot studies exercises were sequenced by clinical case, with several exercises comprising different aspects (for example 3 x elective lymph node CTV exercises, then GTV and finally Local CTV). This allowed interleaving of learning zone themes (i.e. “ABC ABC ABC” practice instead of “AAA BBB CCC”) in keeping with effective instructional design (Rohrer and Pashler, 2010, van Merriënboer and Sweller, 2010).

A written 1-page ‘quick contouring guide’ was provided to trainees (Appendix A.9.3) but not a radiotherapy atlas (i.e. no example contour images were provided).

A post-workshop questionnaire collected trainees’ perceptions of the usefulness (1 = not at all useful, 5 = very useful) of the simulation in general and learning zones in particular. Perceived usability of the simulation was ascertained using a standardised usability instrument - the “system usability scale” (Brooke, 1996). This been validated across a large number of IT systems (Lewis, 2018). The questionnaire also re-checked trainees’ contouring confidence.

Trainees were asked to bring their own laptops. Five extra were kept in reserve; if some trainees had to share this was identified via the post-workshop questionnaire.

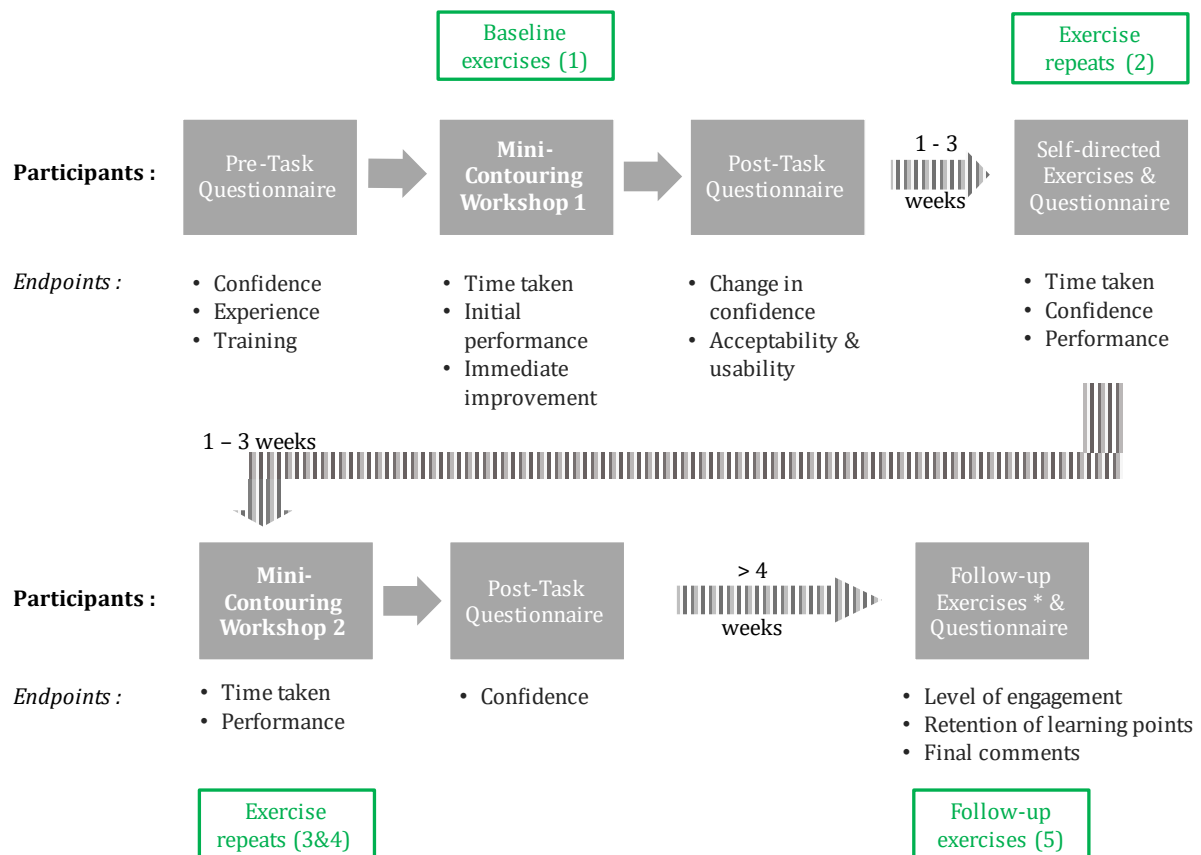
Four weeks after the workshop 4 follow-up contouring exercises were released by email along with a further questionnaire on trainees’ contouring confidence and their reflections on Mini-Contour. After the follow-up exercises were released, trainees were sent up to 2 follow-up e-mails if they had not yet completed them, and provided with a certificate if they had. Trainees were allowed up to 6 weeks (i.e. until 10 weeks after the workshop) to complete the exercises.

9.2.3 Study cohort & procedures - international longitudinal trainee programme

This study was conducted through local and regional training sessions in centres in Canada, the US, and Australia. The allotted time for each live session varied slightly by site but was generally 1 - 1.5 hours.

Study procedures are shown in Figure 9-3. In the first workshop, the facilitators (SLD & local site-specialist faculty) guided trainees through a live tutorial and 8 exercises - these exercises corresponded to the first round of exercises in the UK trainee workshop. Prior lectures and repeat exercises were not possible for these groups due to time constraints. 4 further self-directed

exercises were released by email 4 weeks later: these corresponded to the exercises used as 'immediate repeats' in the UK workshop. Another 1-3 weeks later the second live workshop was conducted, with a further 10 exercises (newly created). Finally 4 follow-up exercises were released >4 weeks after the final workshop:



* Follow-up exercises were the same as for the UK trainee study

Figure 9-3 - Flowchart of international trainee longitudinal programme study procedures

The surveys items were the same, with the exception the addition of 3 more detailed questions specifically about learning zones (see Appendix A.9.4) to the post-task survey, added via a protocol amendment.

Attendees were included if they gave consent and participated in at least the first workshop and pre-workshop survey. During subsequent activities additional attendees were welcomed to participate in the contouring exercises but not included in the study.

As with UK trainees, up to 2 reminder emails for the follow-up exercises were sent with certificates provided on completion. Trainees were allowed up to 6 weeks (i.e. until 10 weeks after workshop 2) to complete the exercises.

9.2.4 Study cohort & procedures - EMBRACE-II

This pilot was conducted during the EMBRACE trial group annual meeting 2019. Study procedures are shown in Figure 9-4. During a 1.5 hour workshop in March 2019 attendees worked through the consent, pre- and post- workshop questionnaires.

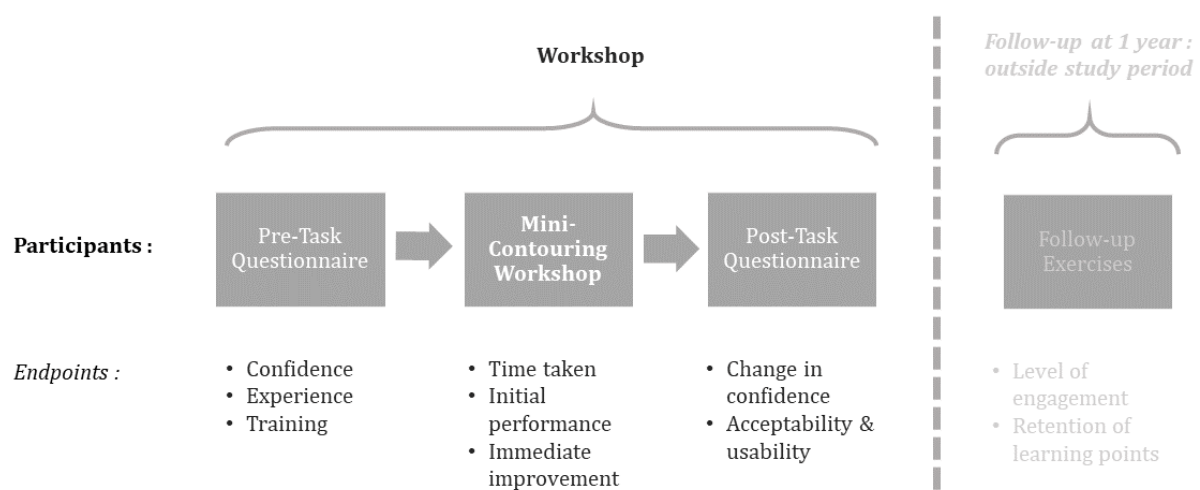


Figure 9-4 - Flowchart of EMBRACE 2019 contouring workshop study procedures

The questionnaires were the same as for trainees with the addition of questions concerning confidence for brachytherapy regions of interest (Appendix A.9.4).

Following the live meeting, EMBRACE group clinicians (a further 86) were invited by e-mail to participate in the contouring exercises during a limited 2-month period. Two further reminder e-mails were sent. The EMBRACE-II trial management group planned follow-up exercises for the following year and so were not performed within this study.

9.2.5 Analysis & statistics

Analysis of quantitative data

Most data were not parametrically distributed. Correlations were calculated using Spearman's rank correlation. Changes in confidence and conformity index performance (paired; see below) were compared using the Wilcoxon signed rank test, and learning zone success rates

(proportions) compared using McNemar's test. These were computed using MATLAB version R2019b (The Mathworks Inc., 2019).

Corrections for multiple testing were not applied, but the occurrence of multiple testing was considered as part of the interpretation of results.

Contour analysis

Jaccard conformity indices ('JCI' see Section 5.2.3) were calculated against all reference contours using MATLAB. In most cases there was more than one reference contour and an average was calculated.

Learning zone success was initially defined as the complete inclusion or exclusion.

Following comments from the UK trainees a post-hoc 'sensitivity' analysis was carried out analysing the effect of inclusion/exclusion 'sensitivity' on learning zone success rates. For inclusion zones sensitivity was defined as the proportion of the learning zone *covered by the user contour*:

$$\frac{\text{User Contour} \cap \text{Learning zone}}{\text{Learning zone}}$$

For exclusion zones sensitivity was defined as the proportion of the learning zone *free from overlap with the user contour*:

$$1 - \frac{\text{User Contour} \cap \text{Learning zone}}{\text{Learning zone}}$$

The latter construct is equivalent to the 'geographic miss index' (Muijs et al., 2009) or 'discordance index' (Hanna et al., 2010); the former can be considered a 'geographic hit index'.

For the EMBRACE-II cohort, the rates of errors seen in the EBRT and brachytherapy accreditation exercises were compared with the percentage of clinicians correctly including or avoiding similarly themed learning zones.

Analysis of qualitative data

Open-ended survey responses were coded by a single researcher (SLD). Codes were grouped into themes (Braun and Clarke, 2006) and the frequency of these were tabulated in each cohort according to the principles of content analysis (Morrison, 2018).

Quotations are presented as per Chapter 8 (see Section 8.2.3).

9.2.6 Ethics

The study had ethical approval from the University of Nottingham (Faculty of Medicine and Health Sciences approval reference: 119-1810) and Health Education England. For all cohorts, local or regional training programme directors gave approval prior to study initiation. All participants gave written informed consent. Attendees who did not give their consent for the study were free to participate in the educational programme including Mini-Contour exercises.

9.3 Results - Exercise creation

Overall time taken and time per stage of preparation for the 16 contouring exercises shared between the two trainee EBRT programmes are presented in Figure 9-5. They took around one hour to prepare on average (mean 63 minutes, range 34 - 81 minutes). Averages are reported as often the time taken for the initial exercise in a case was longer than for subsequent exercise derived from the same case.

The most time-consuming steps were image preparation (average 10 minutes per exercise), image upload & learning zone creation (20 minutes) and exercise review/revision (11 minutes). After review by and discussion with a second faculty member (GE), 8/42 (17%) learning zones required revision. After review by a third faculty member (RN) 5 (12%) of learning zones required revision; 4 of these were the same learning zones revised in the previous review stage, and all were exclusion zones.

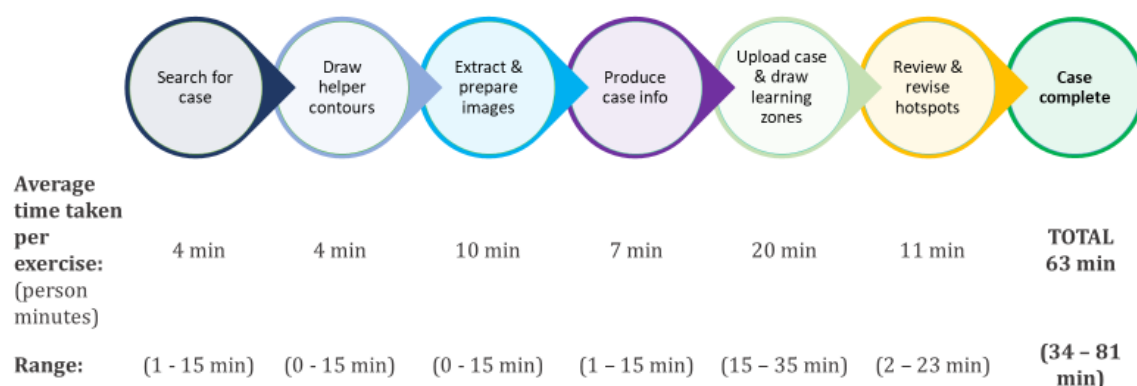


Figure 9-5 - Average time taken for each stage of the Mini-Contour trainee EBRT exercise creation process

The four EMBRACE-II brachytherapy exercises took around 50% longer to prepare - on average 93 minutes - as they required reconstruction of 2 image planes in .jpeg format (Figure 9-6); a process which was only semi-automated. 2 learning zones out of 10 (20%) were revised.

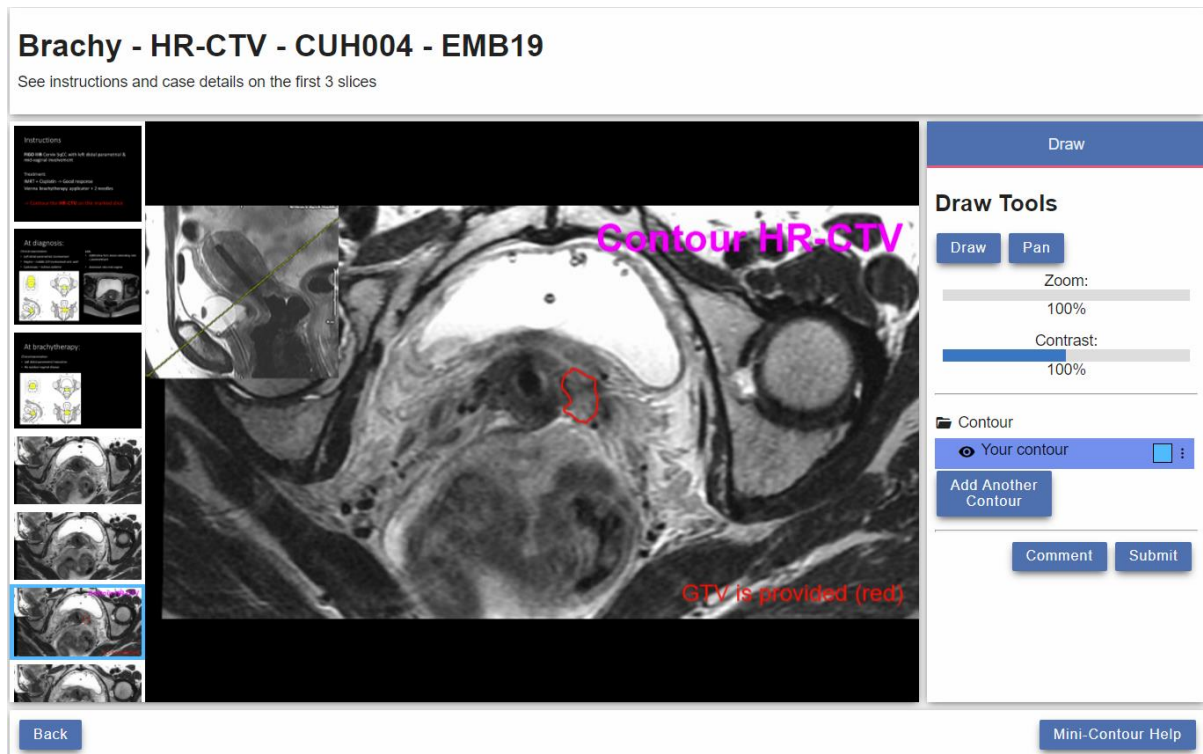


Figure 9-6 - An EMBRACE 2019 contouring workshop brachytherapy Mini-Contour exercise (here the HR-CTV exercise is shown). These took longer to prepare as sagittal slices (see top left of central window) had to be inserted into axial images

9.4 Results - UK workshops

9.4.1 Demographics

In the overall cohort of 80 trainees there was a slight preponderance of trainees in the earlier stages of Clinical Oncology training (Figure 9-7.A). 43 (54%) of the trainees attended the South London workshop, 20 (25%) the North London workshop and 17 (21%) the Manchester workshop.

44 (55%) had some clinical experience of cervix cancer (Figure 9-7.B). Only 6 (8%) had received formal training in cervix cancer contouring outside their own centre (Figure 9-7.C).

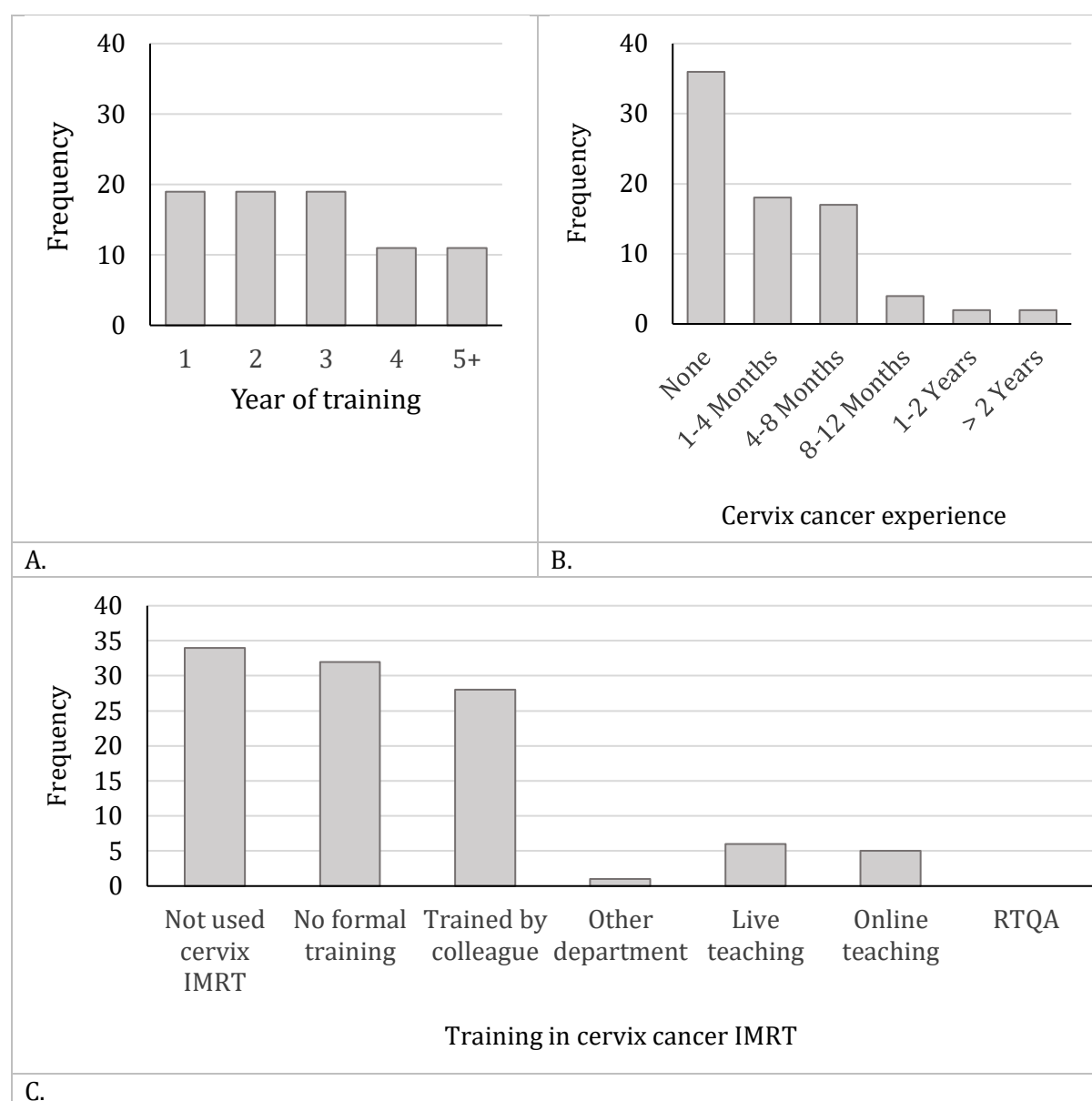


Figure 9-7 - UK trainee participant demographics: stage of training (A), Cervix experience (B) and delineation training (C) of UK trainees

9.4.2 User experience

Timing data

After an initial learning curve of 4-5 exercises, trainees consistently contoured in a median of <3 minutes per exercise (Figure 9-8). This speed was maintained in the follow-up exercises. Every exercise transmitted less than 1MB of data and loaded in less than 5 seconds for all users.

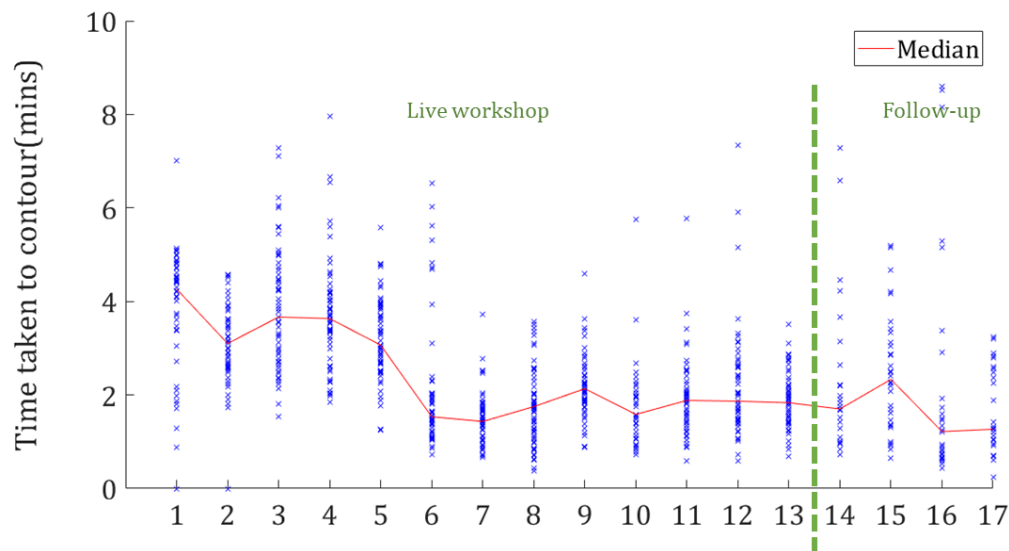


Figure 9-8 - Time taken to contour per exercise for the UK trainee cohort

The average time taken for trainees to complete each case varied significantly, even during the workshop; this spread was wider for the self-directed exercises.

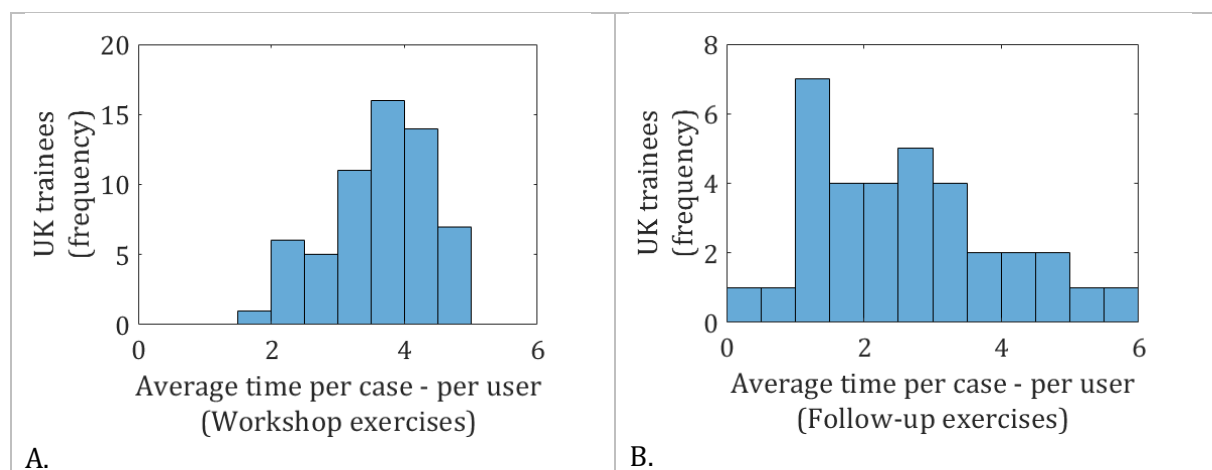


Figure 9-9 - Average time - measured in minutes - taken per case for each UK trainee. (A) Workshop exercises. (B) Self-directed exercises

Usability

76/80 (95%) UK trainees completed the post-workshop survey. Trainees perceived the simulation as easy to use: the median system usability score was 80, compared to a benchmark of 71 across a wide variety of IT systems (Bangor et al., 2008).

UK Trainees generally perceived the simulation to be quite similar to contouring in clinical practice (median = 80% 'similar to contouring in real life', interquartile range 70-90%, range 32-100%).

Usefulness

UK trainees generally enjoyed learning using the simulation and found it useful: 73/76 (96%) trainees responding to the post-workshop survey agreed or strongly agreed with the statement “The Mini-Contour tool is useful for learning radiotherapy delineation” and 71/76 (93%) with “I enjoyed learning using the Mini-Contour tool”. 74/76 (97%) respondents agreed or strongly agreed that they would be interested in future delineation practice using the simulation.

Trainees reported that they found the automated learning zone feedback very useful (median score 5/5 = “very useful”, mean 4.5, range 3-5), and that they generally agreed with the learning zones’ locations and feedback (median score 5/5 = “strongly agree”, mean 4.3, range 3-5). Only one trainee used the *inbuilt* functionality to disagree with a learning zone (excluding the bladder from the local CTV), citing erroneous learning zone assessment: “I included and it said I didn't!”. Trainees did however submit free-text comments in the post-workshop survey about the high stringency of learning zones (see below).

Qualitative comments - content analysis

Appendix Table A.9-2 shows the full results of the content analysis of UK trainees’ free-text comments and suggestions. Many included general positive comments (n=27), for example: “Very useful learning tool and would recommend it for anyone learning how to contour”.

Suggestions to expand the content to other tumour sites (n=10) and/or add more exercises (n=5) were common: “A very good start to solving a very important learning need! Useful. Thank you. But you would need many different images to practice on including different stage of tumour [sic] (and different tumour types of course) etc. to become comfortable contouring in general..”.

The most common suggestion for improving Mini-Contour was to upgrade the contouring functionality (n=29): most commonly trainees requested a ‘rollerball’/‘brush’: “Joining dots technique is not so user friendly and our contouring software is more advanced. It would be useful to be able to use a brush function?” and “... a rollerball and eraser would make it a lot easier to use. I started to get repetitive strain injury with my fingers from having to click with points.”. Several users (n=4) commented that the contrast function did not work for them.

Several (n=7) trainees commented specifically that the learning zone assessment was overly stringent, for example: “it will tell you [that] you are wrong when you have 'touched a line'. This does not mean you have missed and so will frustrate juniors who may lose confidence if they consistently get negative feedback ... [...] ... Perhaps there should be an amber area for

discrepancy of a mm or 2. As in true to life planning, you will see these kind of marginal discrepancies between all operators”.

Trainees also commented that they envisioned this could be used early in the process of their clinical attachment, for example: “Really good to have for each specialty and use prior to starting the job to learn the basics”.

Other noteworthy comments included the value of working through at one’s own pace, the difficulty in starting these exercises without any clinical experience, and insight regarding skill loss in-between the workshop and testing: “I think I left it too long after teaching to do follow up but would be useful in a tumour site I was more familiar with”.

9.4.3 Confidence & performance

Of the 80 trainees, 28 shared computers and submitted contours jointly, therefore a maximum of 66 workshop submissions for each exercise were available for analysis. Shared submissions (n=14) were excluded when assessing the relationship between initial confidence and performance.

In the pre-workshop survey trainees were generally most confident in delineating in organs at risk, and less confident in delineating the GTV, CTV and ITV:

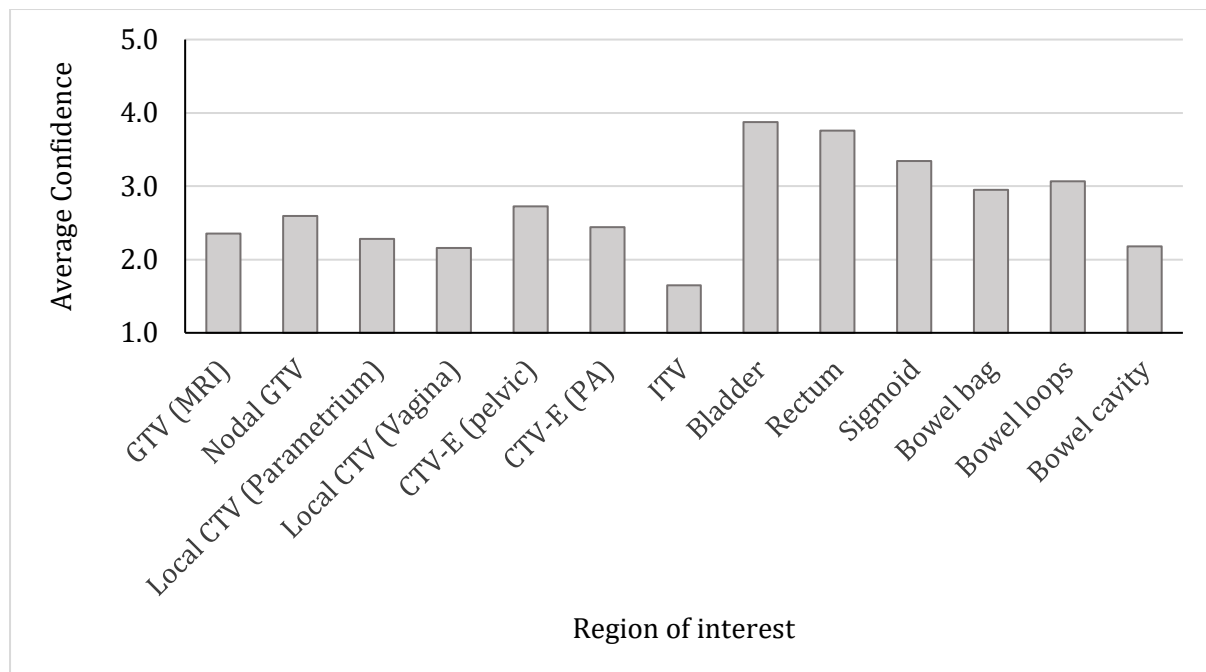


Figure 9-10 - UK trainees' average pre-workshop confidence per cervix cancer EBRT region of interest

For the pelvic lymph node CTV, increased confidence in contouring correlated strongly with a higher stage of training (Spearman's rank correlation $\rho = 0.74$ (pelvic - see Figure 9-11) & 0.69 (para-aortic), $p < 0.01$ for both).

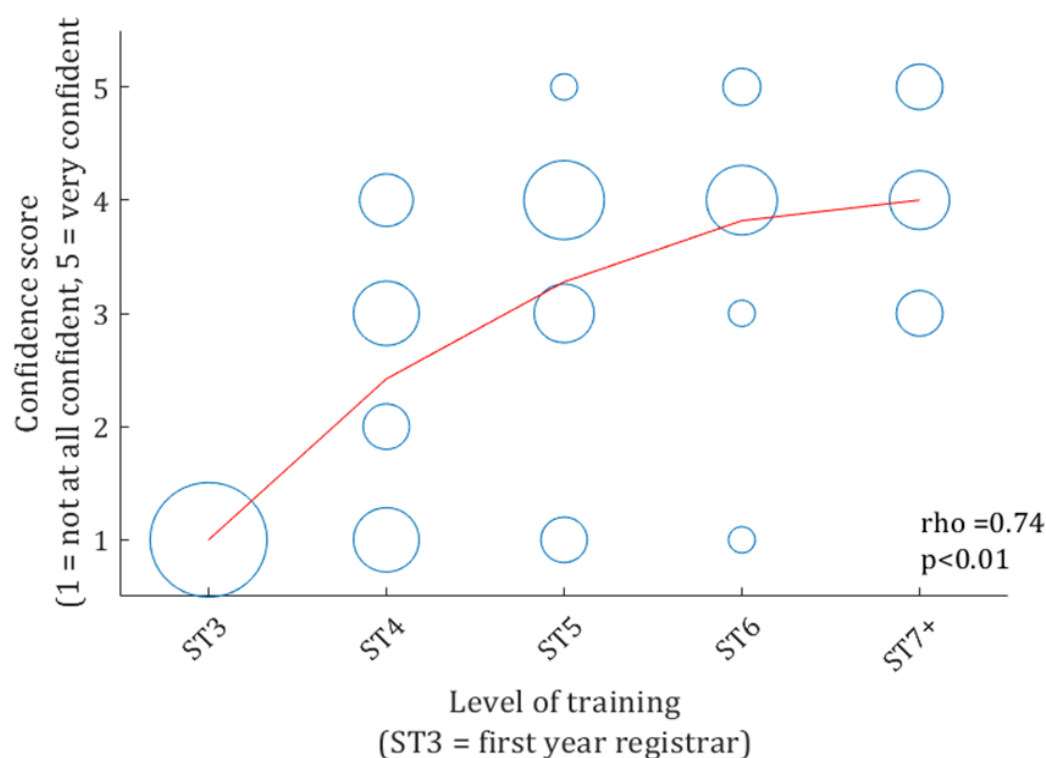


Figure 9-11 - Bubble chart comparing UK trainees' confidence in contouring the pelvic lymph node CTV with their stage of training. Each bubble's size is proportional to the number of trainees. The red line represents the average confidence for each group.

There was no correlation between pre-workshop confidence and initial performance, whether performance was assessed by learning zones or JCI (Table 9-2). The overall correlation between trainees' cervix cancer experience and their performance was also generally weak or negligible and/or not statistically significant (Appendix Table A.9-3).

Table 9-2 - UK Trainees' pre-workshop average confidence and correlation of individual confidence with ranked performance on the first relevant learning exercise ("N/S" = $p > 0.05$)

Region of interest	Mean baseline confidence (/5)	Correlation with initial performance by learning zones (Spearman's rho)	Correlation with initial performance by Jaccard conformity index (Spearman's rho)
GTV (on MRI)	2.4	-0.02 (N/S)	0.11 (N/S)
Local CTV - parametrium	2.4	0.17 (N/S)	0.15 (N/S)
Elective lymph node CTV - pelvic	2.7	0.25 (N/S)	0.26 (N/S)
Elective lymph node CTV - para-aortic	2.4	0.24 (N/S)	0.07 (N/S)

The median and spread of JCI values for trainees' contours was variable across the exercises (Figure 9-12).

Median JCI values generally increased when a particular target volume was repeated within the workshop, but did not always (e.g. GTV; exercises 1 & 2). There were failed inclusion zones for some trainees whose contours had a very high JCI - for example 27/34 (79%) of trainees who missed the left lateral para-aortic lymph nodes in the first exercise still had a JCI >0.8.

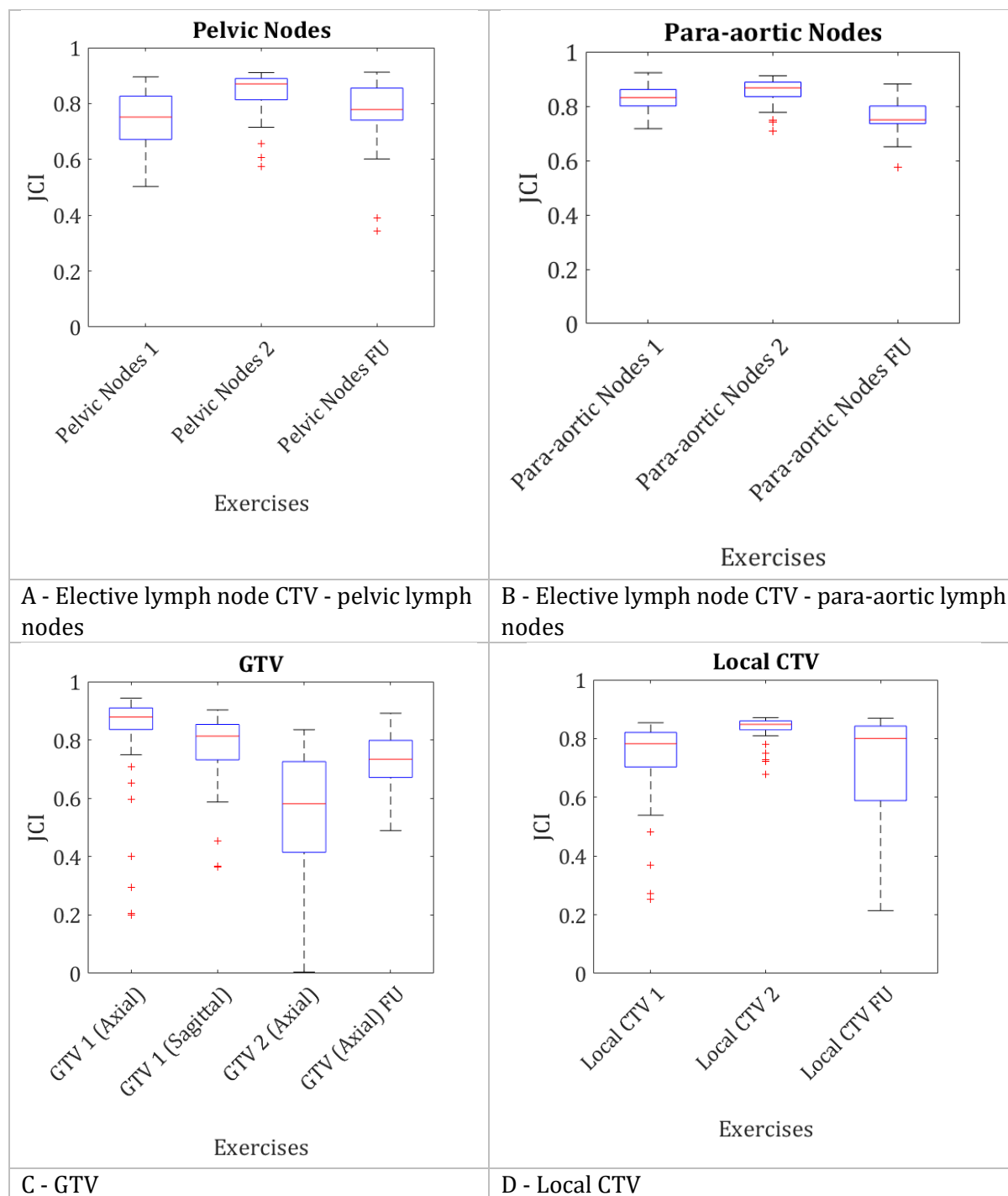


Figure 9-12 - Performance of UK trainees by Jaccard conformity Index (JCI) across repeated exercise themes: A - Elective CTV - pelvic lymph nodes. B - Elective CTV - para-aortic lymph nodes. C - GTV. D - Local CTV . FU = "follow-up".

For the 3 inclusion zones (i.e. targets) that were repeated during the workshop, there was an improvement in success rates from 40-48% to 80-92% ($p < 0.01$) on immediate re-testing (Figure 9-13). For the 5 'exclusion' learning zones that were repeated, 2 consistently had success rates $> 80\%$. For the other 3, performance slightly decreased (N/S).

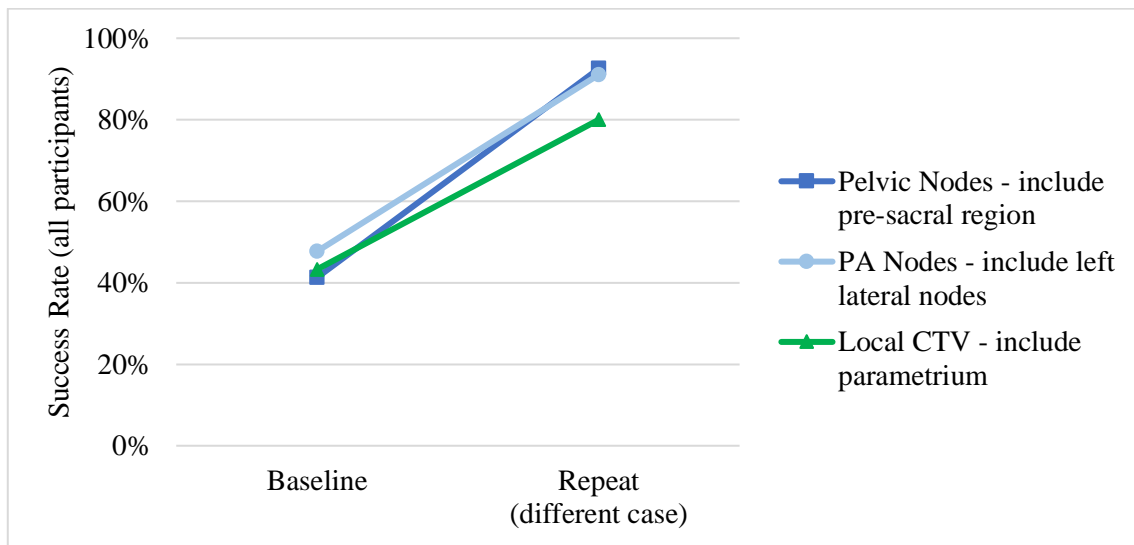


Figure 9-13 - UK trainees' performance for **include** learning zones during live workshop

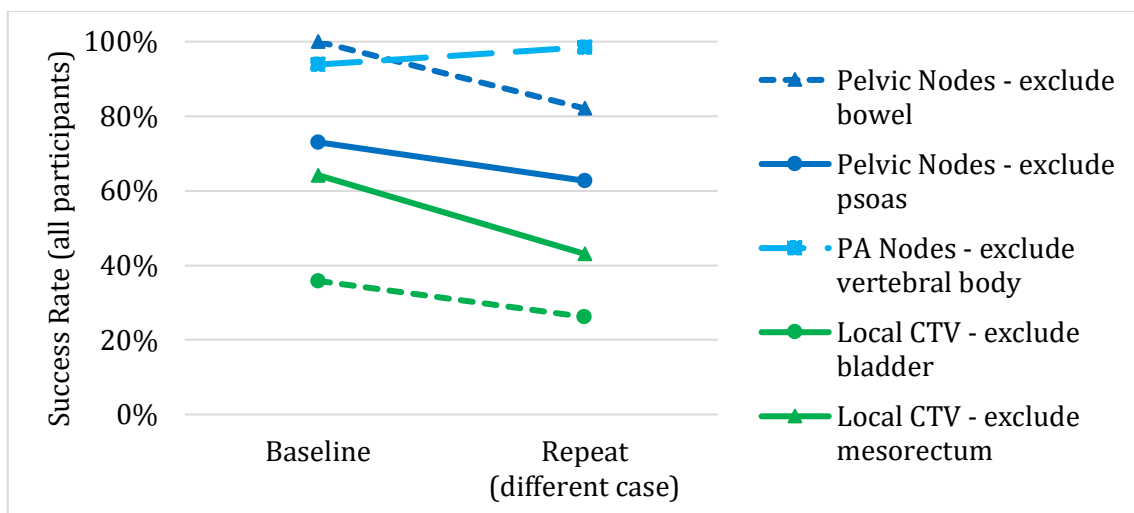


Figure 9-14 - UK trainees' performance for **exclude** learning zones during live workshop

After the workshop, trainees' contouring confidence had increased by 1.4/5 points ($p < 0.01$) on average across the target volumes tested (GTV, Local CTV : Parametrium, CTV-E pelvic & para-aortic nodes). Confidence in contouring the target volumes which were **not** practiced (Lymph node GTV, Local CTV: vaginal region, and ITV) also increased, by 1.3 points on average ($p = 0.6$ for practiced vs un-practiced target volumes).

9.4.4 Follow-up exercises

34/80 (43%) trainees completed the follow-up exercises, a median of 6 weeks after the live workshops. Of these, 30/34 (88%) completed the follow-up survey. Confidence in delineation generally remained elevated when compared to baseline (pre-workshop) levels (Figure 9-15.A). For the learning zones that were re-tested, performance on all three 'inclusion' zones fell; for 2 out of 3 to near or below baseline (Figure 9-15.B). There was no relationship between trainees' stage of training and their skill retention for inclusion zones at follow-up. For exclusion zones, follow-up performance was highly variable (Figure 9-16).

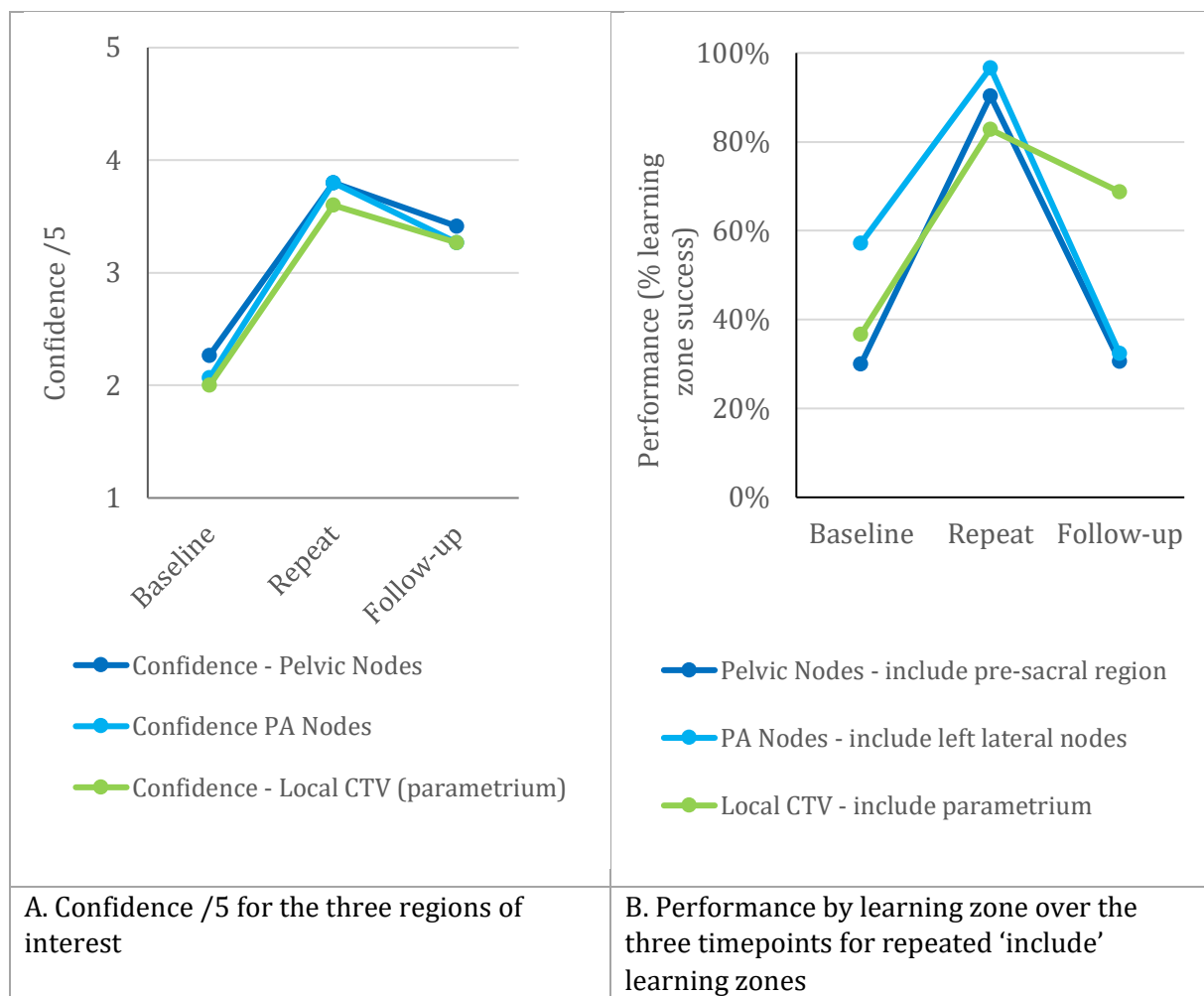


Figure 9-15 - Performance & confidence on the **inclusion** zones & related regions of interest for trainees who completed the follow-up exercises. Only trainees who completed all 3 timepoints are included.

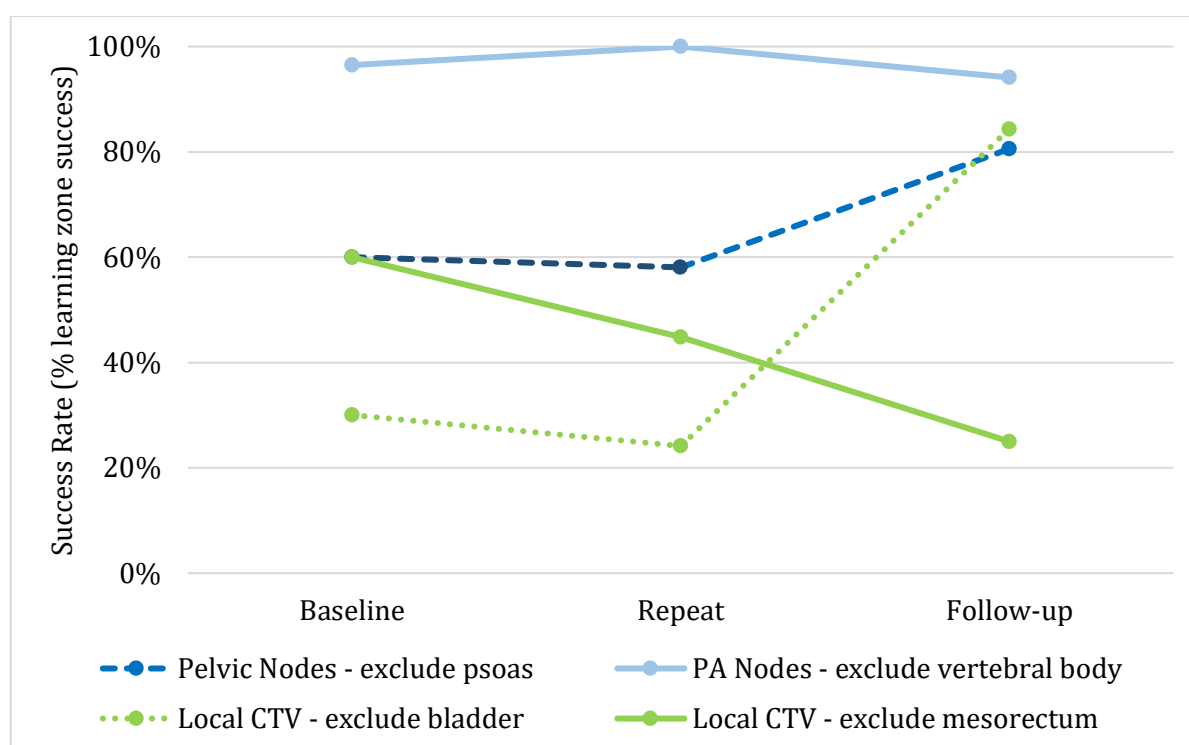


Figure 9-16 - Performance & confidence on the **exclusion** zones & related regions of interest for trainees who completed the follow-up exercises. Only trainees who completed all 3 timepoints are included.

9.4.5 Learning zone sensitivity analysis

Following the UK trainees' comments regarding the sensitivity of learning zones to very small infractions, a post-hoc analysis of the effects of learning zone stringency on the 'learning zone performance' endpoint was carried out.

Appendix Table A.9-4 details the effects of adjusting the learning zone leniency, i.e. the proportion of 'acceptable' overlap between a learner contour and any learning zone, from 80% to 100% for include learning zones and from 20% to 0% for exclude learning zones (i.e. from significant leniency to maximum stringency. Maximum stringency was the user experience).

Figure 9-17 demonstrates the effects of varying the learning zone leniency on four example learning zones. The shape of the include learning zone success rate graphs generally remained similar with variations in leniency. The 'exclusion' learning zones were sometimes much more sensitive to small variations in leniency (see Figure 9-17 & Appendix Table A.9-4). In three exercises increasing the acceptable overlap with an exclusion zone from 0% to 2% resulted in a >50% increase in participant success rate, dramatically changing the shape of the graph (e.g. Figure 9-17.D).

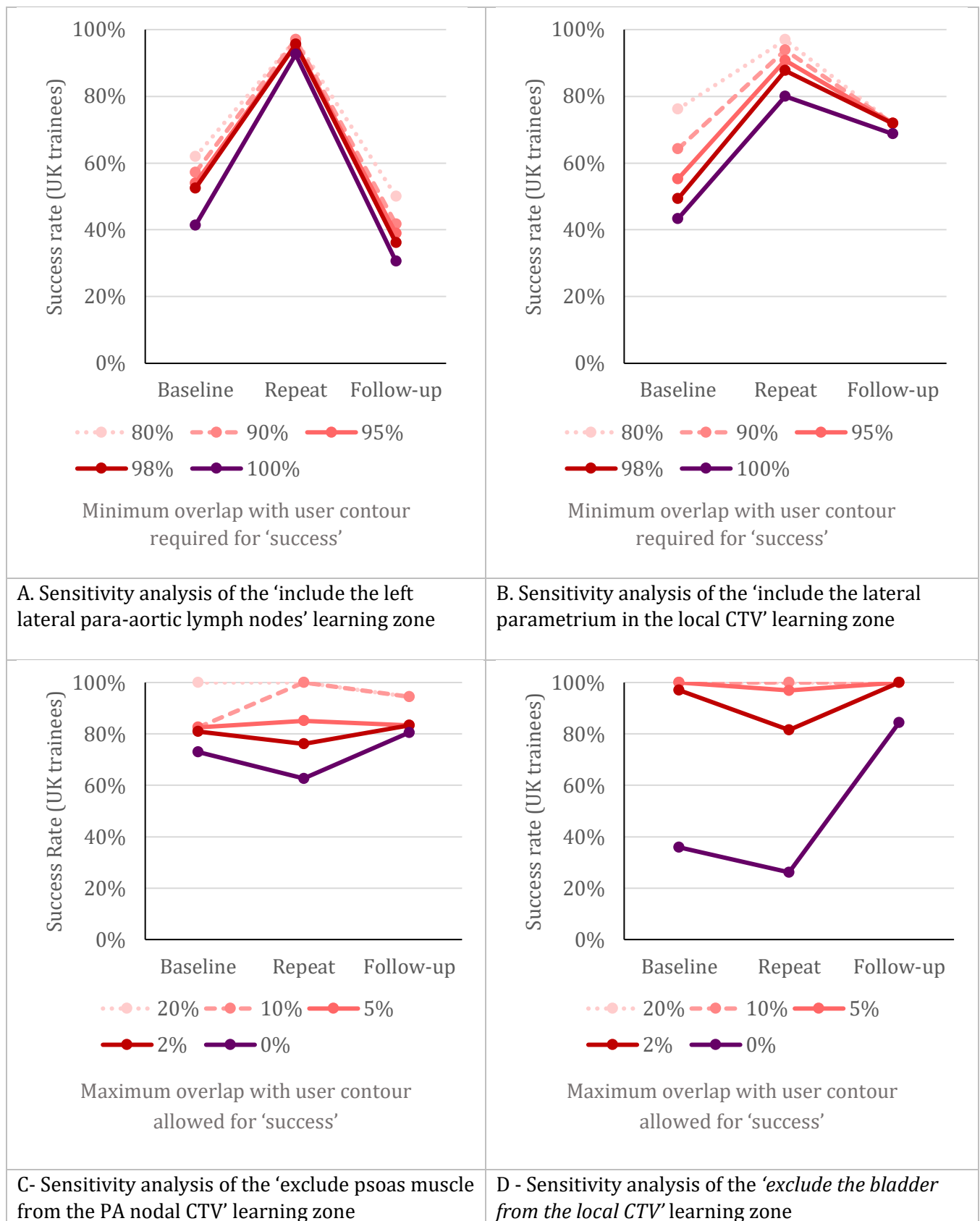
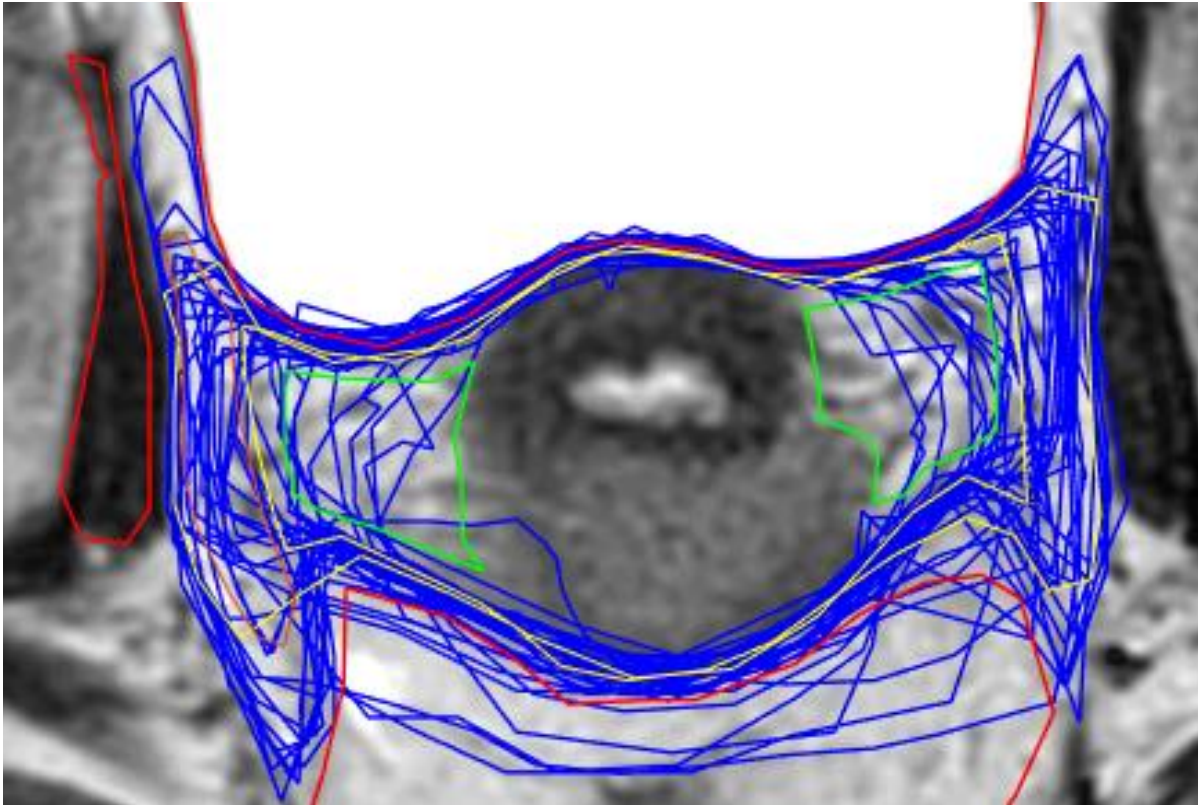


Figure 9-17 - Analysis of the sensitivity of selected learning zones to variations in stringency in the UK trainees cohort. A & B: two include learning zones; C&D: two exclude learning zones.

This sensitivity of the 'exclude bladder' learning zone is illustrated with the user contours in Figure 9-18 below: all of these users would have passed the learning zone if the stringency was reduced to allow a small (2%) overlap:



*Figure 9-18 - UK trainee contours that failed the '**exclude bladder**' learning zone that would have passed that particular learning zone if 2% overlap was allowed*

9.5 Results - international trainee longitudinal programme

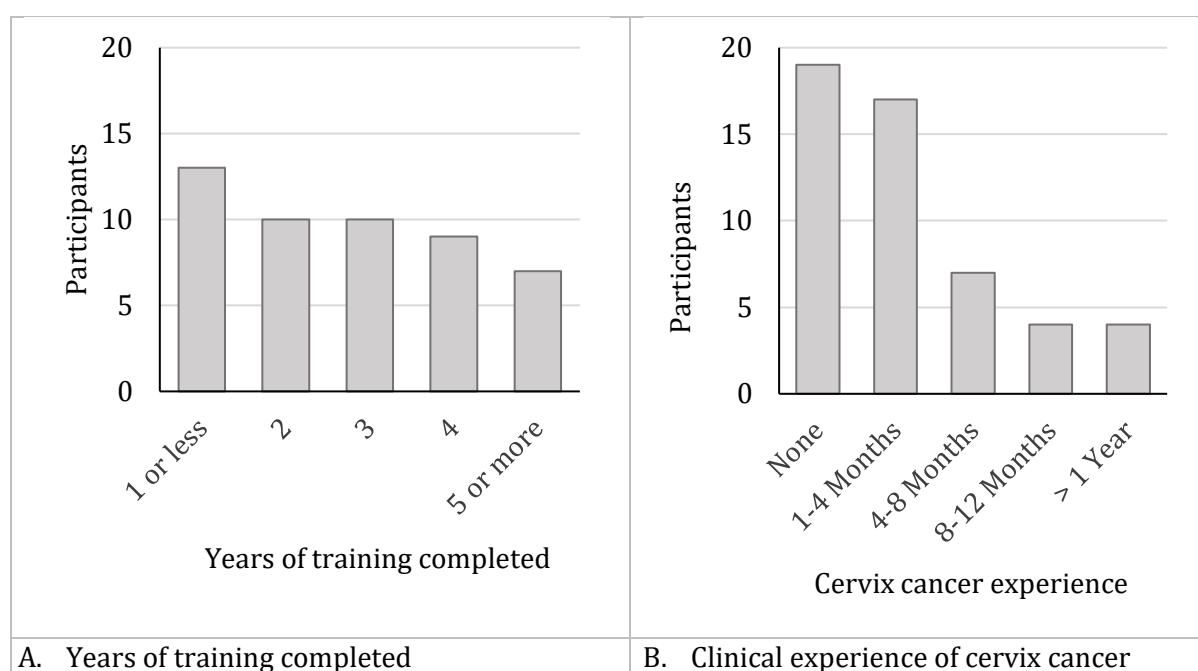
9.5.1 Demographics & engagement

51 participants enrolled from five centres across North America and Australia, in groups of 9 to 12 trainees in each location (Table 9-3).

Table 9-3 - Location and timing of international trainees participating in longitudinal study

Location	n	Study timeframe
Princess Margaret Hospital, Toronto, Canada	9	May - September 2019
McGill University Health Centre, Montreal, Canada	11	June - October 2019
Cross Cancer Institute, Edmonton, Canada	9	July - November 2019
New South Wales training scheme, Australia	12	September 2019 - January 2020
MD Anderson Cancer Centre, Houston, USA	10	November 2019 - January 2020

As with the UK trainees there was a slight preponderance of earlier stages of training (Figure 9-19.A). This cohort were very slightly more experienced in cervix cancer than the UK cohort: 32 (63%) had previous clinical experience (Figure 9-19.B) and 8 (16%) had received formal training in cervix cancer IMRT outside of their own centre (Figure 9-19.C; c.f. Figure 9-7.B & C).



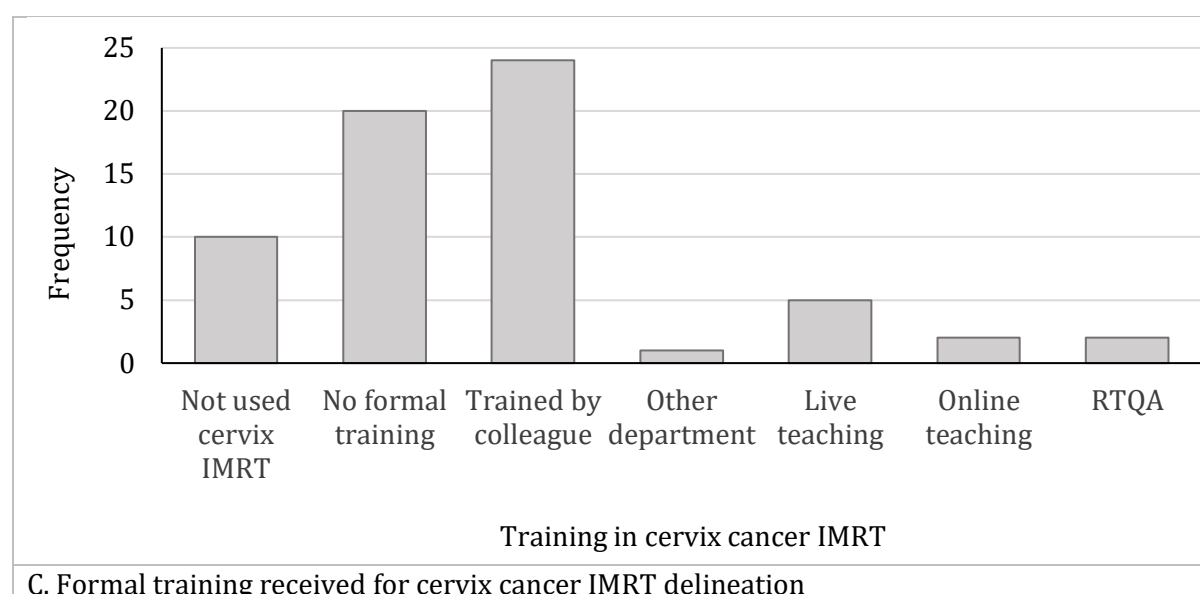


Figure 9-19 - International trainee longitudinal participant demographics: stage of training (A), cervix cancer experience (B), and training in cervix cancer delineation (C)

There was a downwards trend in engagement over time and fewer participants completed the self-directed modules than the live modules (Figure 9-20).

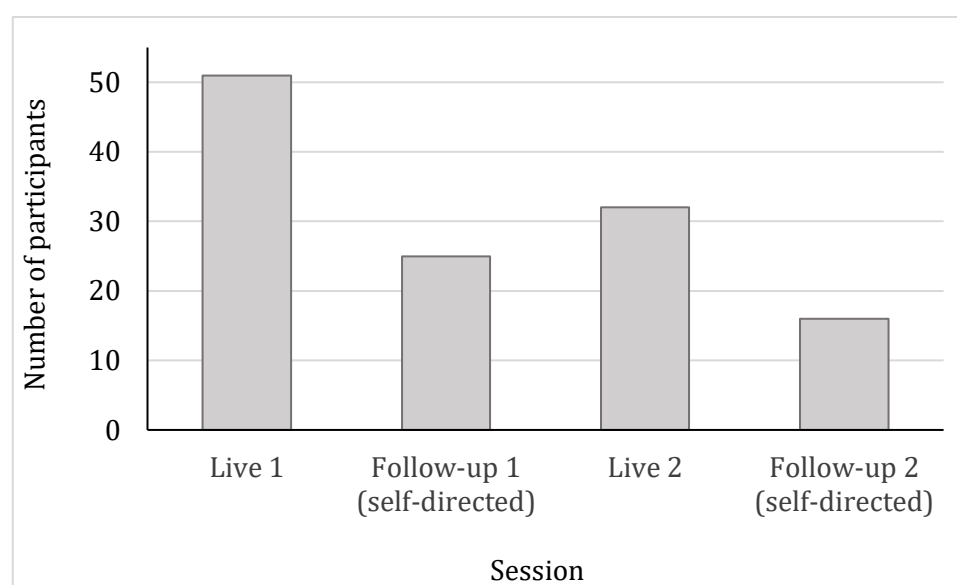


Figure 9-20 - Number of participants completing each session of longitudinal Mini-Contour training programme

Only 7/51 (14%) participants engaged with all four modules as designed. A further 4/51 (8%) completed all but the first self-directed module. For convenience, these 11 trainees will hereafter be referred to as the 'engaged' trainees. Participation in the follow-up exercises was 33% (17/51) - slightly lower than the UK trainees.

9.5.2 User experience

Timing data

The trend in the time taken for trainees to contour an exercise was similar to the UK workshops (Figure 9-21). There was a small uptick in contouring time at the start of each batch of exercises.

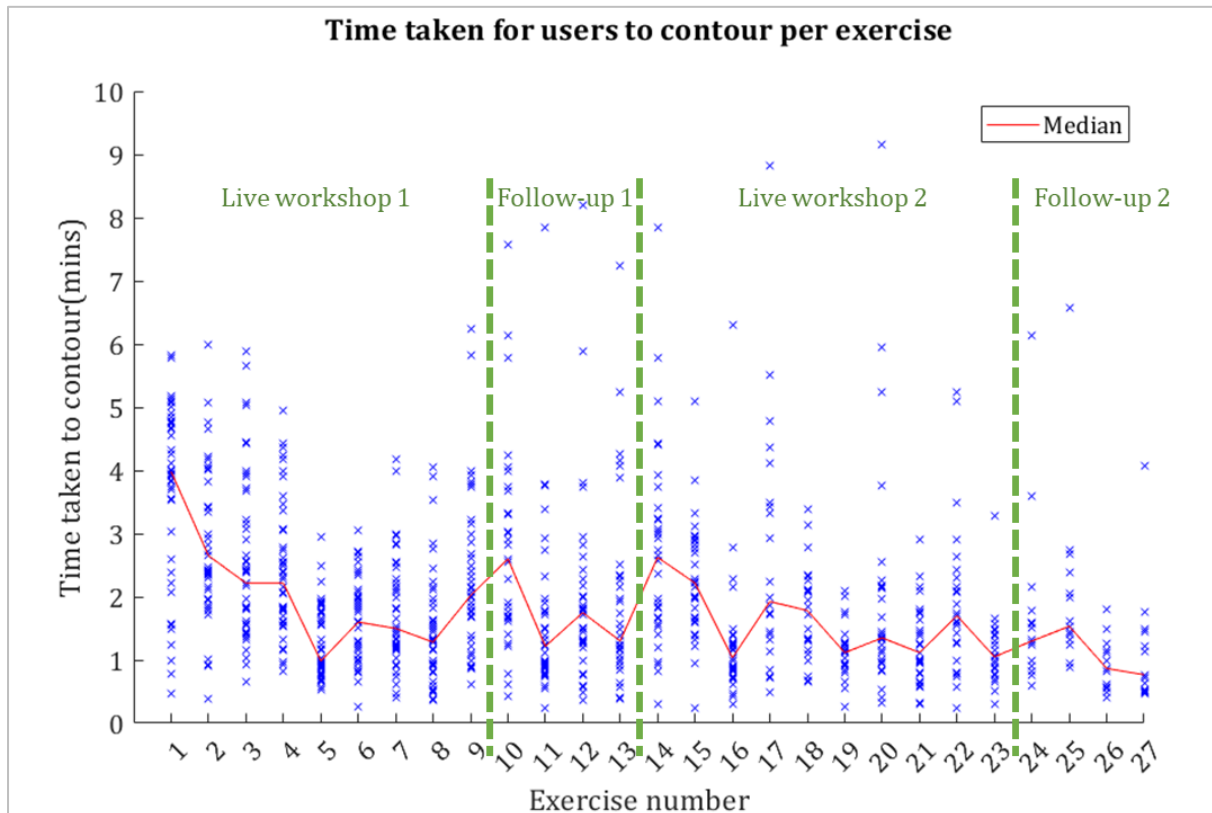


Figure 9-21 - Time taken to contour per exercise for longitudinal trainee contouring programme

In the live workshop the median time taken to review the reference contour and learning zone feedback was 2.7 minutes (average 2.9) - see Figure 9-22 and Figure 9-23.A:

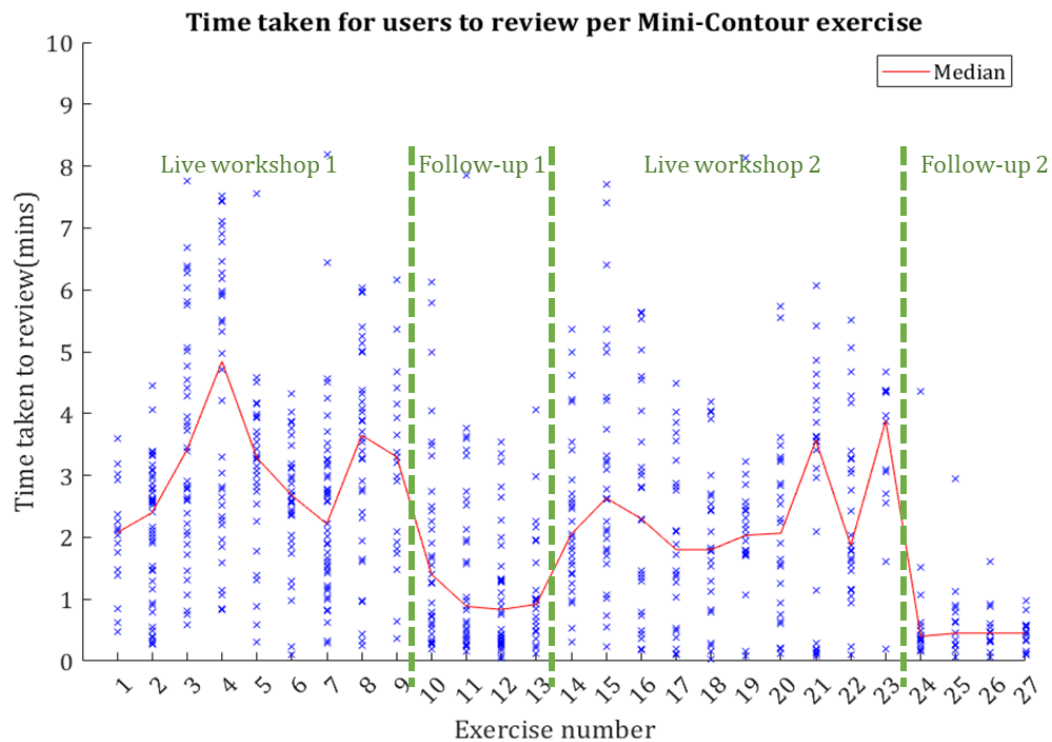


Figure 9-22 - Time taken to review feedback per exercise for longitudinal trainee contouring programme

For the self-directed exercises trainees typically took around 2 minutes less time to review the exercise after submission (median 0.47 minutes, average 0.85 minutes, $p < 0.001$, Wilcoxon signed rank test); Figure 9-23.B shows a histogram of time taken for self-directed review; in the majority of instances (51%) trainees took less than 30 seconds to review the feedback. In 7% of instances trainees took less than 10 seconds.

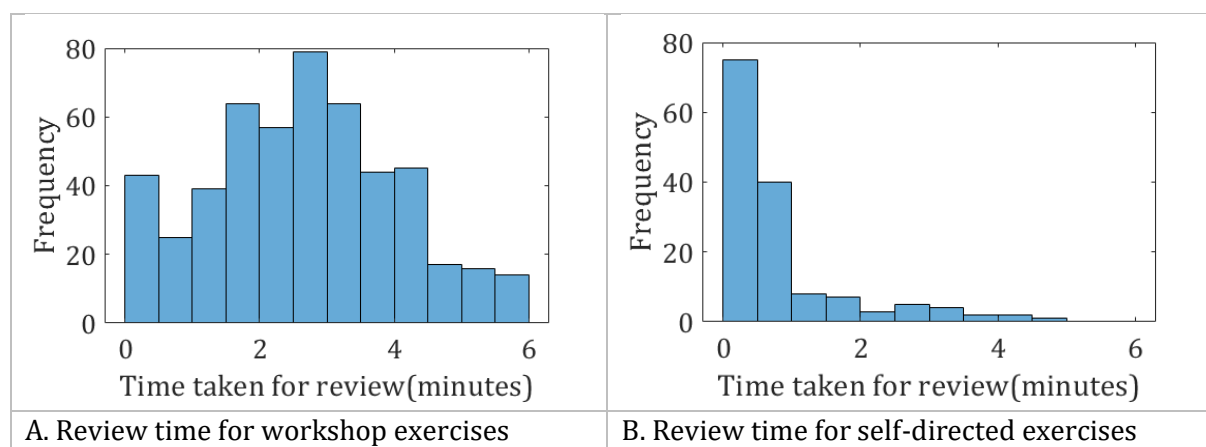


Figure 9-23 - Histogram of time taken for international trainees to review feedback in: (A) workshop and (B) self-directed exercises

Usability

49/51 (96%) international trainees completed the post-workshop survey. The median system usability scale score was 78, similar to the UK trainee cohort.

The international trainees generally perceived Mini-Contour to be quite similar to real life (median 78%; inter-quartile range 70-86; range 27-100). In their free-text comments (n=23) international trainees most commonly noted that mini-contour lacked specific drawing/editing tool functionality (n=7), for example: “although clicking to define an outline is certainly possible I rarely use it due to curved surfaces and do most contours with the brush tool with auto fill-in”) and that exercises either lacked 3-D viewing or mandated single-slice contouring (n=6) - for example: “not on multiple axial slices (full volumetric 3D contours are important for GYN)”.

Usefulness

International trainees reported that they generally found the learning zone feedback very useful (median = 5/5 [“very useful”], mean 4.5, range 3-5). 94% agreed or strongly agreed that they would be interested in future delineation practice with Mini-Contour.

25 (49%) trainees gave an example of when they learned from learning zone feedback, most commonly: extension of the elective lymph node volume into the left para-aortic space (n=8); parametrial borders (3); and paying attention to clinical examination findings (3).

18 trainees gave examples of when they disagreed with learning zone placement and/or feedback; for 14 of these this related to the stringency of assessment after being marked incorrect with only slight overlap with an ‘exclude’ learning zone, for example: “the smallest sliver of mesorectum was included in the volume, so I did not think it should be considered as wrong” - illustrated in Figure 9-24:

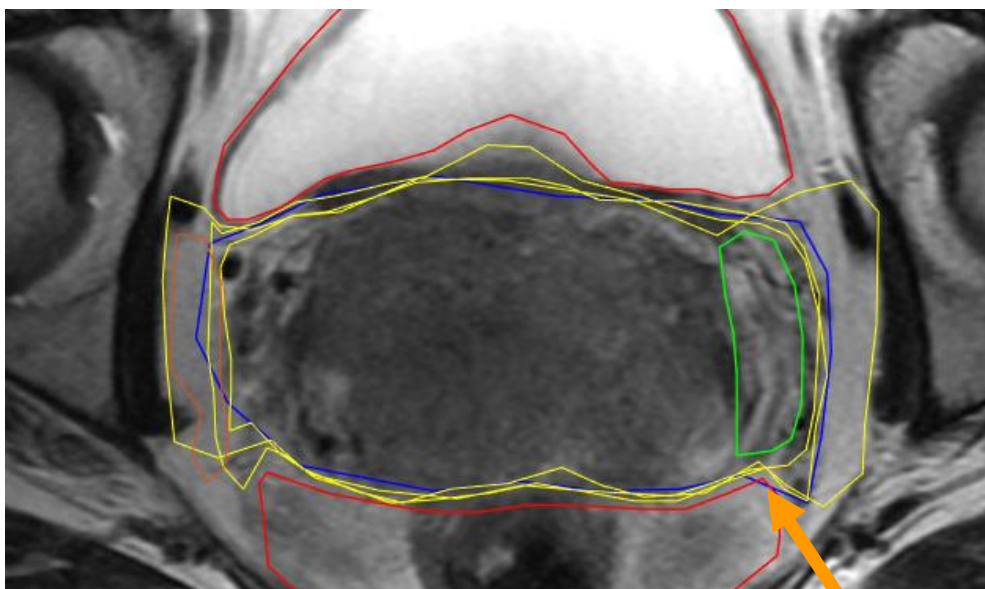


Figure 9-24 - A trainee contour (in blue) resulting in a comment regarding overly stringent assessment. Orange arrow indicates area of overlap with 'exclusion' learning zone.

Two trainees remarked specifically that they had been penalised for what they considered acceptable contouring variation - interestingly for one of these it concerned the lateral parametrial border which in fact was not 'assessed', only illustrated with a comment (i.e. a neutral remark). No trainees used the inbuilt Mini-Contour functionality to dispute or comment on learning zones despite encouragement to do so during the workshop sessions.

International trainees gave 35 examples during the programme of when they had learned something from a feedback - the most common example was the lateral para-aortic lymph node CTV learning zone (n=12), but also several others (see Appendix Table A.9-5).

Other qualitative comments

When trainees were asked about their thoughts in general (open-ended), there were many positive comments. Most were non-specific (n=24); the most common features specifically praised were the automated feedback (8), for example: "... especially like the real time feedback and ability to toggle over the anatomy to see why you would include or not-include something in your target volume", and referred to the simplicity (4) and psychological safety (4) of the simulation: "the anonymous contour (so we are not embarrassed) was excellent".

Suggestions for improvements were fewer than in UK trainees (n=6) and consisted of: improving the drawing interface, reducing the learning zone stringency, 3-D imaging, anatomy pre-learning for junior learners, and partnering with other organisations already running contouring programmes.

9.5.3 Confidence & performance

Confidence after the initial workshop increased for the relevant target volumes (by, on average, 1.2 points - Figure 9-25), but also for the target volumes not practiced and the organs at risk as seen with the UK trainees.

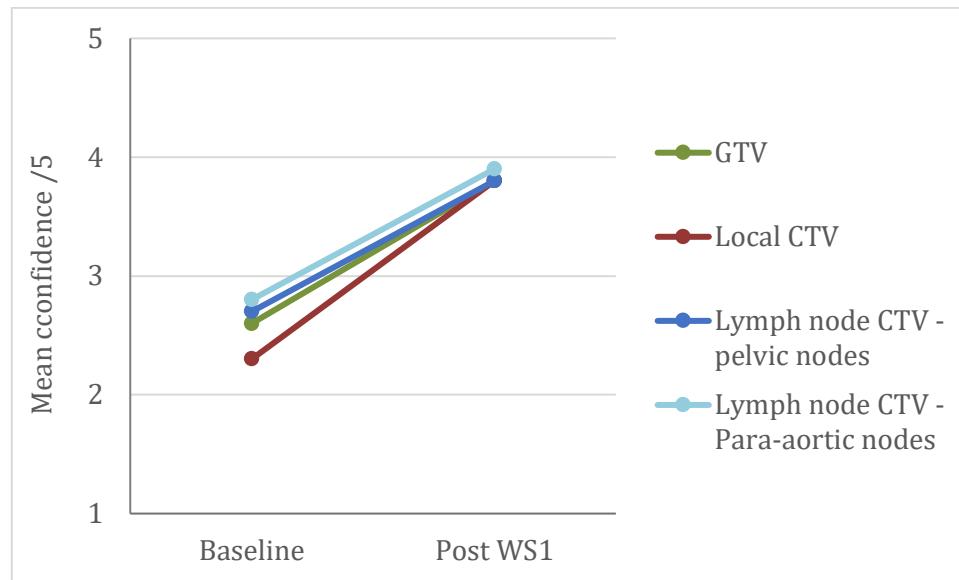


Figure 9-25 - Confidence at baseline and after the first workshop for all trainees in the longitudinal programme

Changes in confidence over time in the engaged group are shown in Figure 9-26; after the initial increase in confidence there generally was a slight decrease for all target volumes after the second workshop and at follow-up.

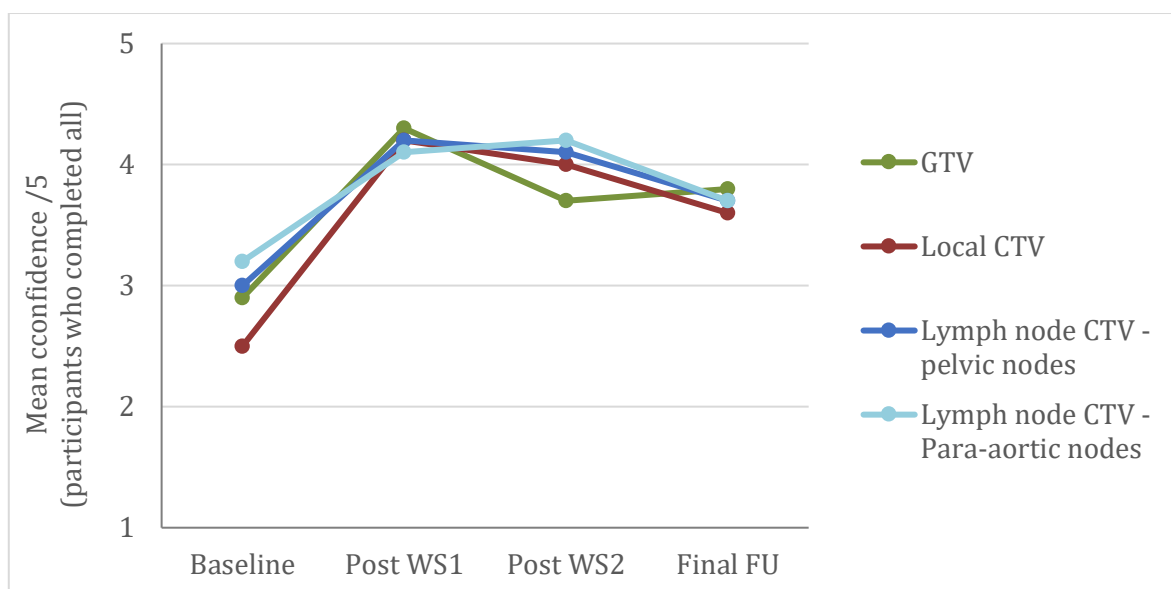


Figure 9-26 - Confidence for 4 regions over time for the 11 most engaged trainees in the longitudinal programme

As with the UK trainees, the correlation between confidence and performance - whether assessed by conformity index or learning zones - was generally weak or negligible, and not statistically significant (Table 9-4). Correlation of performance with international trainees' cervix cancer experience was statistically significant for both learning zone and JCI performance, but negligible or weak in strength ($\rho = 0.13$ & 0.22 respectively).

Table 9-4 - Relationship between confidence and performance for international trainees

Region	Assessment	ρ	p
GTV	Learning zone score	0.29	N/S
	Jaccard conformity index	0.25	N/S
Local CTV	Learning zone score	-0.05	N/S
	Jaccard conformity index	0.23	N/S
Lymph node CTV - pelvic nodes	Learning zone score	0.27	N/S
	Jaccard conformity index	0.36	0.02
Lymph node CTV - Para-aortic nodes	Learning zone score	0.23	N/S
	Jaccard conformity index	0.17	N/S

Trends in performance over time during repeats of the same exercise theme are shown for all submissions, regardless of engagement, in Figure 9-27 (JCI) and Figure 9-28 (learning zones). Learning zone performance over time of only the engaged participants is shown in Appendix Figure A.9-2 - the shapes of the graphs are similar.

Boxplots of conformity indices per exercise theme repeat are shown in Figure 9-27. When comparing along the course of the longitudinal programme conformity for the pelvic and para-aortic lymph nodes was higher in the final compared to the first exercise (0.66 vs 0.84; 0.70 vs

0.80, $p < 0.01$ for both) was lower in the final exercise than the first exercise for 2/4 exercise themes - this was statistically significant for the GTV (0.85 vs 0.79, $p < 0.01$) exercises. The shape of the conformity index plots were similar for the 11 engaged trainees (Appendix Figure A.9-1).

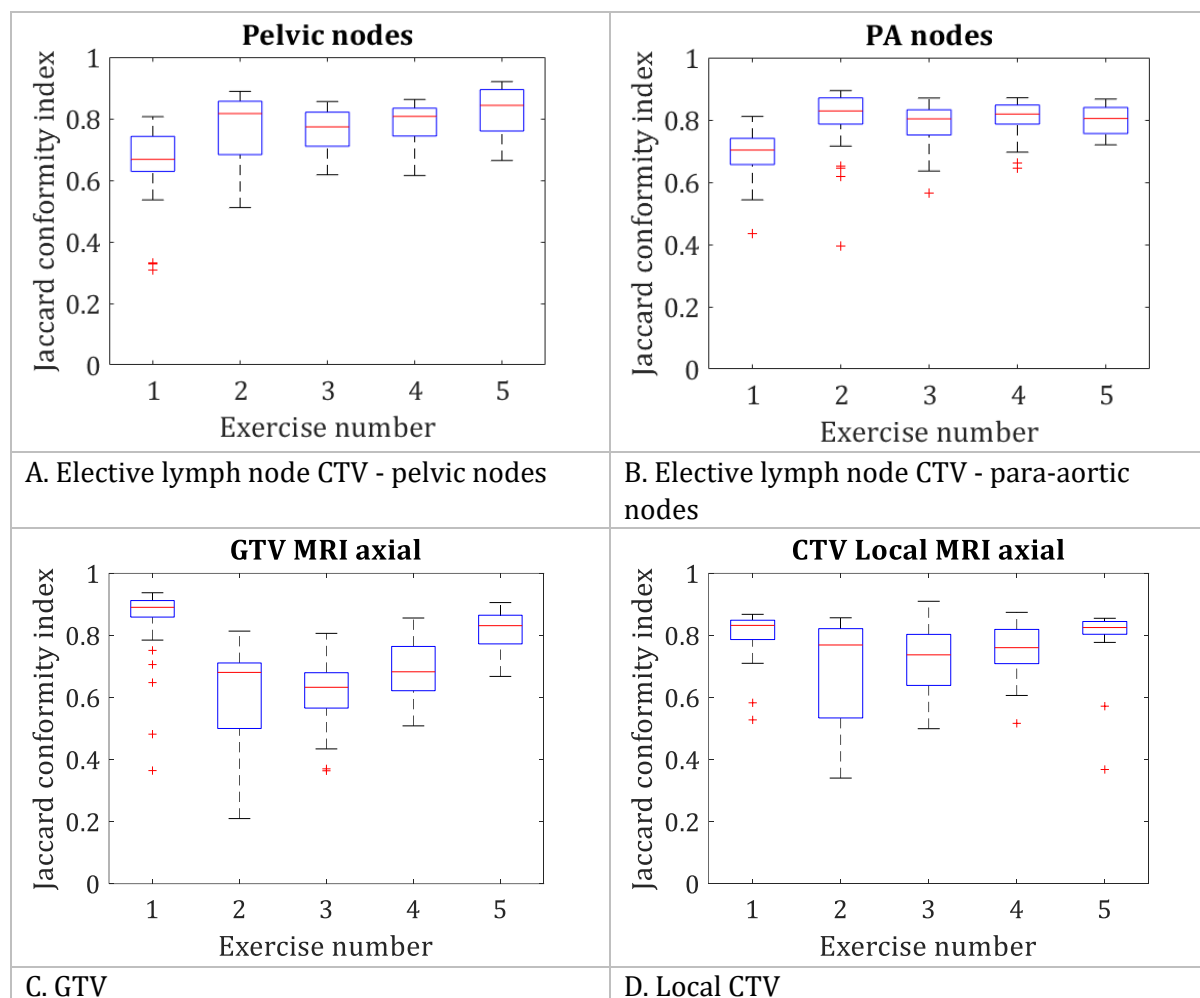


Figure 9-27 - Conformity indices for repeated exercises over the course of the international longitudinal programme - all participants

Baseline performance assessed by conformity index was lower compared to the UK trainees for the elective lymph node volumes, but higher for the local CTV:

Table 9-5 - Mean conformity indices at baseline, 1st repeat and follow-up for the UK and international trainees. The **same** exercises are compared between the two programmes. JCI = Jaccard conformity index

Exercise theme	Cohort	Baseline median JCI		Repeat median JCI		Follow-up median JCI	
Elective lymph node CTV - pelvic	UK	0.75	} p<0.001	0.87	} p<0.01	0.78	} N/S
	International	0.66		0.83		0.84	
Elective lymph node CTV - PA*	UK	0.83	} p<0.001	0.87	} p<0.001	0.75	} N/S
	International	0.70		0.82		0.80	
GTV	UK	0.88	} p<0.001	0.81	} p = 0.047 **	0.73	} p = 0.02 **
	International	0.84		0.77		0.83	
Local CTV	UK	0.78	} p<0.001	0.85	} p<0.001	0.80	} N/S
	International	0.89		0.72		0.82	

* PA = "para-aortic"; ** Questionable statistical significance given multiple comparisons

When comparing the progress of the international trainees and the UK trainees over the same exercises (Table 9-5), the longitudinal cohort had somewhat better conformity in the follow-up in those exercises despite starting at a lower baseline and having less improvement on repeat exercises, but these differences were not statistically significant due to the low numbers of trainees completing follow-up. An exception to this was that international trainees' conformity in the follow-up exercise slightly reduced for the local CTV, as opposed to the UK trainees whose conformity slightly increased compared to baseline.

For include learning zones, i.e. tumour-related or lymph node targets, no definite pattern of substantial improvement was seen over the 5 repeats (Figure 9-28.A). Neither reducing assessment stringency nor analysing the engaged trainees separately substantially altered this finding. Baseline performance was considerably higher for international than for the UK trainees in one inclusion zone (the parametrium).

Performance by learning zone for 2/3 'include' regions was slightly higher (12 - 20%) for the final exercises than the first exercises, although these differences were not statistically significant. This was in contrast to the UK follow-up performance which was essentially the same as at baseline. For the local CTV ('include parametrium') learning zone follow-up performance was slightly lower than a high baseline. As seen in the UK workshops, performance across the 'exclude' learning zone regions was variable (Figure 9-28.B).

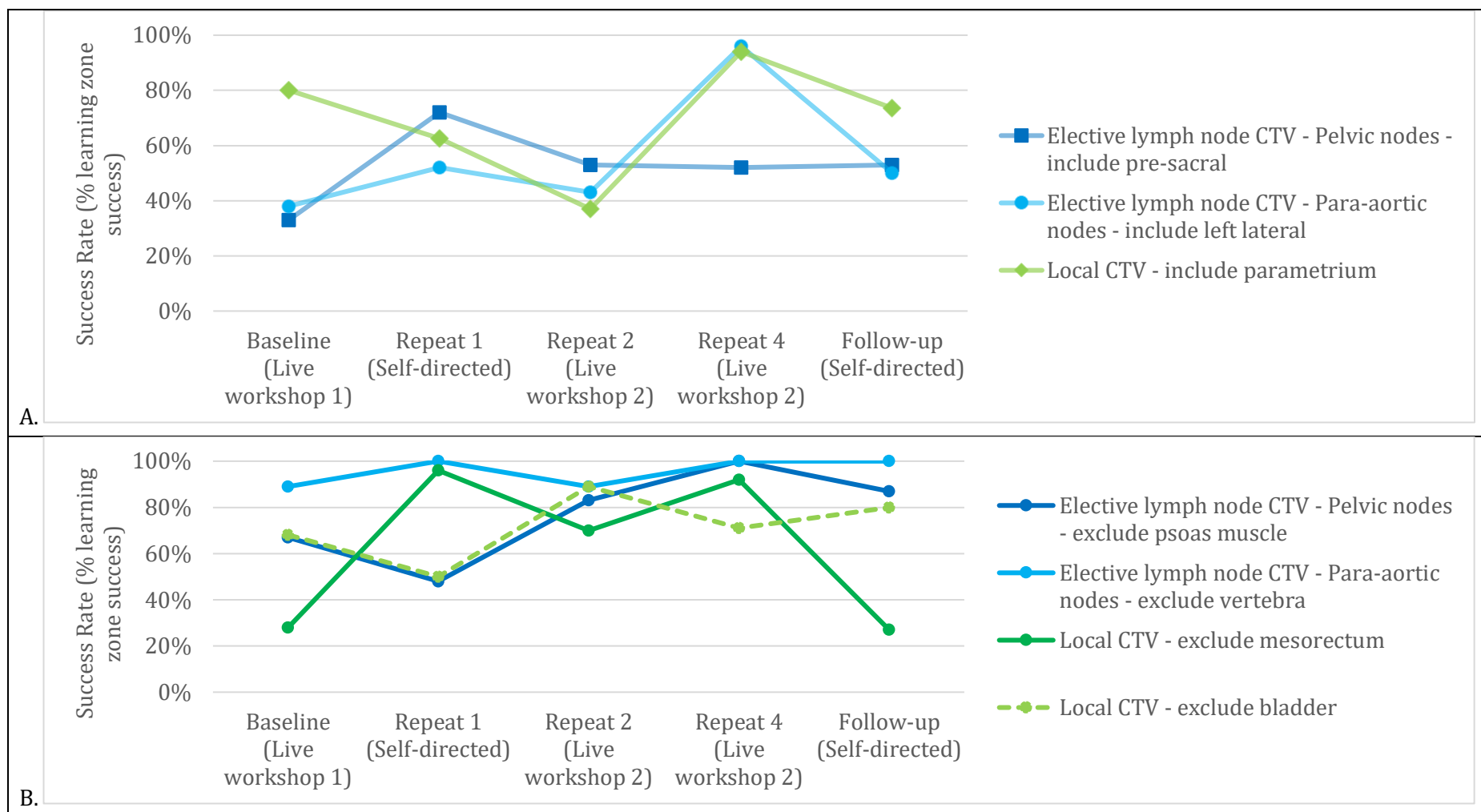


Figure 9-28 - International trainees' performance on 'include' (A) and 'exclude' (B) learning zones repeated over the course of the programme - for all participants

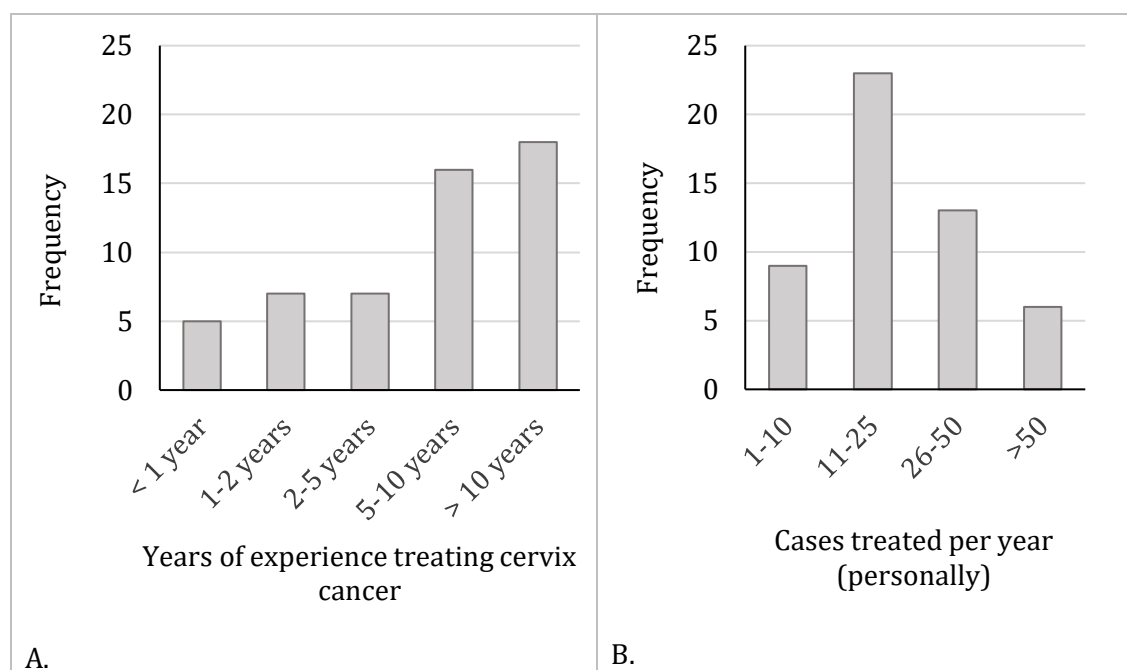
9.6 Results - EMBRACE-II contouring workshop

9.6.1 Demographics & engagement

32 accredited clinicians took part in the EMBRACE-II Annual Meeting workshop in March 2019 and enrolled in the study. Seven other clinicians present did not bring laptops and so were not included in the study. A further 30 enrolled and participated in the following 4 week period after receiving the email invitation. Overall 62 clinicians out of a total of 118 eligible (non-faculty) clinicians in the trial group actively participated (53%). 26/62 (42%) were principal investigators ("PIs") and 36 (58%) were "non-PIs".

56/62 clinicians (90%) completed the pre-workshop survey. 12 (19%) had used Mini-Contour before.

Two thirds had more than 5 years' experience of treating cervix cancer (Figure 9-29.A).



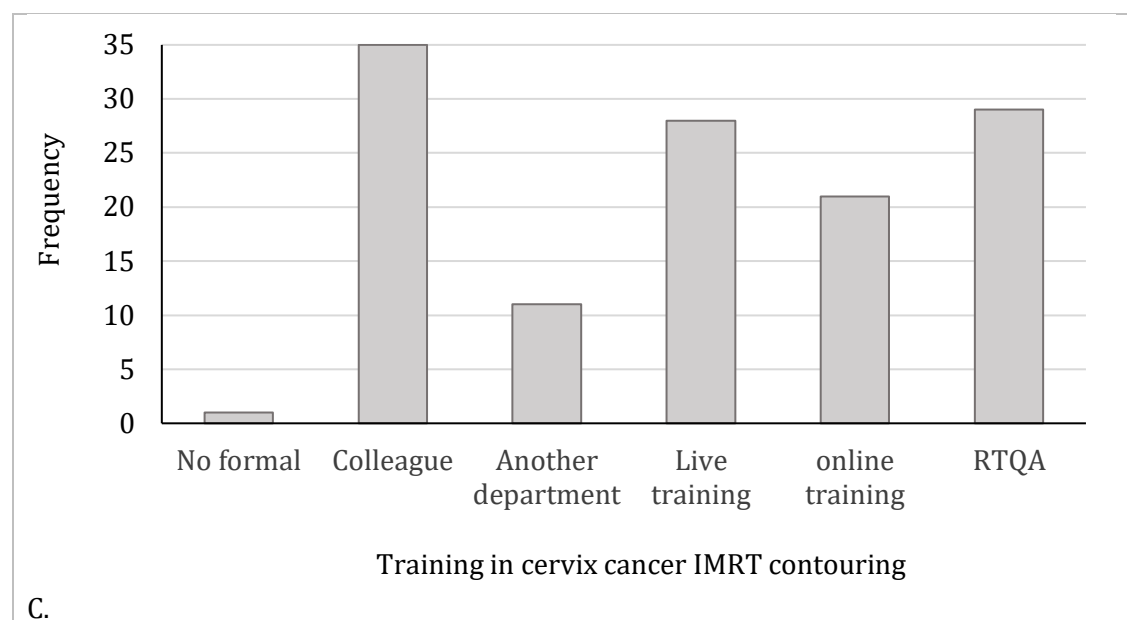


Figure 9-29 - EMBRACE group 2019 annual meeting workshop: participant experience and training in cervix cancer IMRT

Substantially more had attended training courses than in the trainee cohorts (Figure 9-29.C; c.f. Figure 9-7 & Figure 9-19) - 46 (74%) had participated in formal training in cervix cancer delineation.

Clinicians completed a median of 8 out of 9 planned exercises. Fewer self-directed clinicians completed all the exercises (16/30 = 53%) than those attending the live workshop (25/32 = 78%).

9.6.2 User experience

Timing data

Time taken for EMBRACE-II clinicians to contour each exercise is shown in Figure 9-30. The median was consistently below 3 minutes.

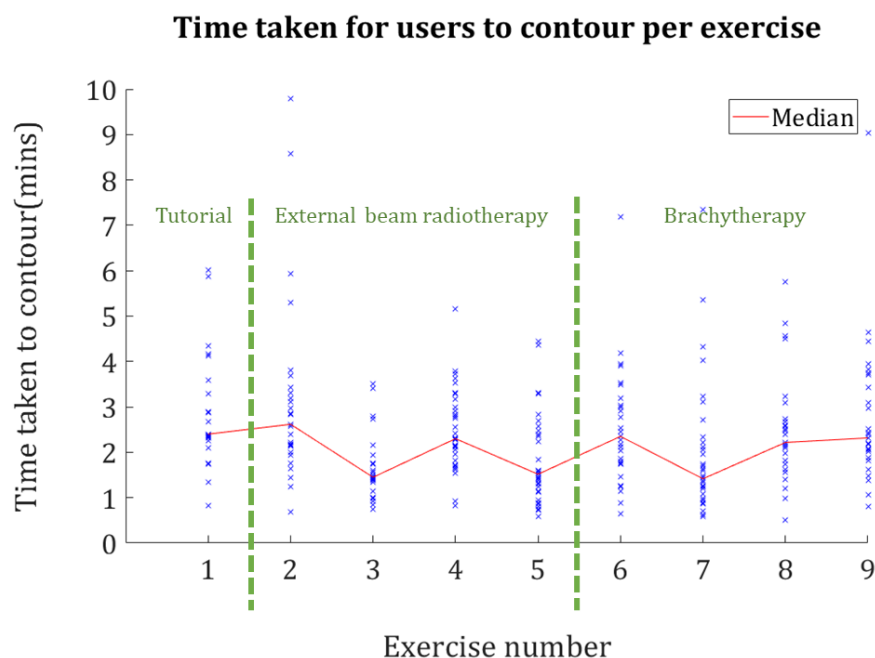


Figure 9-30 - Time taken for EMBRACE-II clinicians to contour per exercise: includes both self-directed and live participants

Time taken to contour and review exercises for self-directed EMBRACE clinicians are displayed in Figure 9-31; again, the data were right-skewed. The median time taken for self-directed review was 1.0 minutes; double that of self-directed trainees in the longitudinal study (trainee median = 0.47, $p < 0.001$; EMBRACE average = 1.7 minutes, c.f. 0.85 minutes for trainees). Only in 1% of instances did clinicians take less than 10 seconds to review the feedback, compared with 7% of instances in the international trainee self-directed cohort.

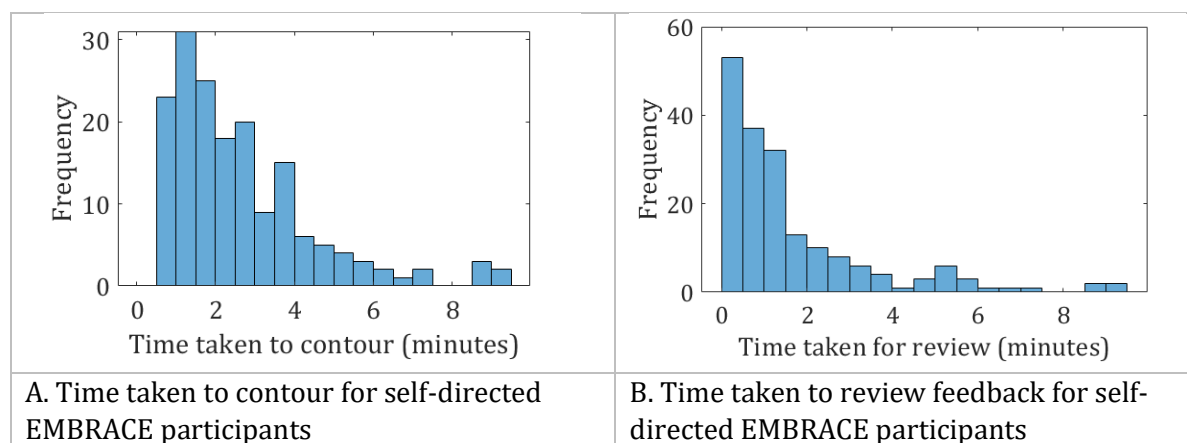


Figure 9-31 - Histograms of time taken for self-directed EMBRACE participants to submit a contour (A) and review feedback (B)

The average time taken to complete each exercise varied considerably for self-directed clinicians: the fastest was 0.78 minutes per exercise and the slowest 5.9 minutes (median 3.6 minutes).

Usability

52/62 (84%) clinicians completed the post-workshop survey. The median system usability scale score was 85 (IQR 75-93; range 43-100). The scores were slightly higher than the UK and international trainees' scores (medians 78 and 80) but these differences were not statistically significant on Kruskal-Wallis analysis ($p=0.08$). The median score was similar (86.5) when only including the 80% of clinicians who had not used Mini-Contour before.

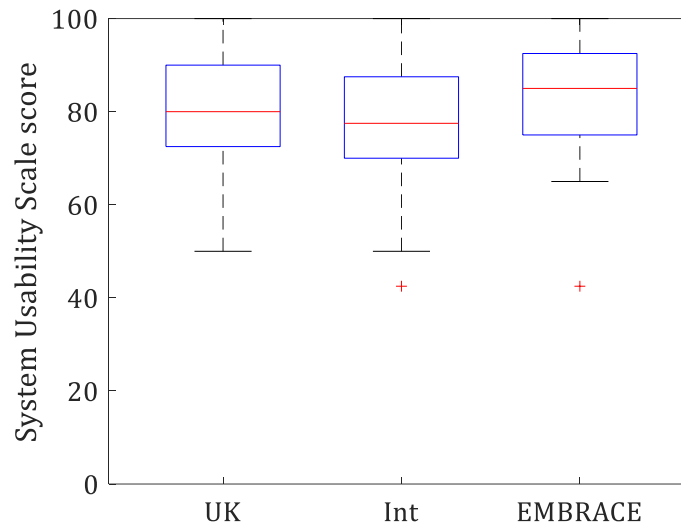


Figure 9-32 - Boxplots showing System Usability Scale scores across all three cohorts

The median reported 'similarity to real life' (i.e. fidelity) was 78% (IQR 70 - 86; range 27-100): very similar to the median percentages reported by UK trainees (80%) and international trainees (78%).

In their free-text comments about fidelity ($n=26$ - see Appendix Table A.9-6) clinicians most commonly noted that: Mini-Contour lacked drawing and/or editing functionality ($n=8$), 3-D imaging (6) or viewing functionality (5), and gave limited clinical information (3) - for example:

"The information you get is limited, you can't use axial, coronal and sagittal planes at the same time, you can only see limited number of slices. It's very difficult to correct the contours"

"I'm more used to contouring using a paint brush tool. Also, I view my contours 3-dimensionally (not just in 1 plane) when contouring. For these 2 reasons (and jet lag), my contouring performance in this exercise was suboptimal".

Clinicians also noted that the way Mini-Contour was used could potentially bypass or exacerbate issues with usability or fidelity, for example:

“Most treatment planning softwares [sic] offer additional tools like brush, free drawing pencil etc.,. Since we are better accustomed to using those tools, some may find it difficult initially to contour using the point-pencil. Nevertheless, it is not a difficult one to master with practice, and I believe inconsistencies will be within acceptable limits.”

“If good representative slices are chosen then it [percentage similarity to real life] is 90. If poorly representative slices are chosen then it is a 10”.

Usefulness

Most EMBRACE-II clinicians reported finding the learning zone feedback useful or very useful (median response = 5/5 [“very useful”], mean 4.6, range 3-5). They reported that they generally agreed with the location and feedback for the learning zones (median response = 4/5 [“agree”], mean 4.2, range 2-5). 92% (48/52) of responding clinicians agreed or strongly agreed that they would be interested in future delineation practice with Mini-Contour (median response = 5/5, mean 4.5, range 1-5); 1 felt neutrally, 1 strongly disagreed and 2 did not answer that question.

17 respondents (33%) gave an example of when they learned from learning zone feedback, most commonly on the brachytherapy intermediate-risk CTV (n=7) and the EBRT local clinical target volume (n=5).

14 respondents (27%) gave examples in the survey of when they disagreed with learning zone placement and/or feedback. A further 10 (16%) clinicians logged 23 distinct learning zone comments within Mini-Contour using the inbuilt functionality. Most commonly these related to the stringency of automated assessment, or to incorrect placement of the learning zone - for example: “The ‘learning zone’ didn’t recognise that I had included whole Cx [cervix] in HR-CTV, and fed back that it should be included, when it clearly was” and “vessel included in HR-CTV” (Figure 9-33). This learning zone was subsequently adjusted.

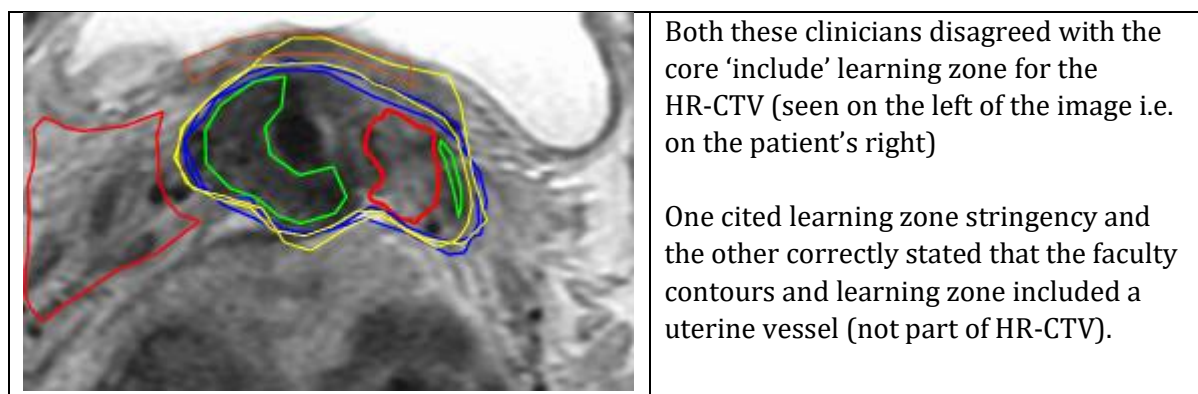


Figure 9-33 - Contours of two EMBRACE-II clinicians (blue) who disagreed with a HR-CTV learning zone

Conversely, clinicians occasionally also logged their disagreement when making a conceptual error (n=4), despite learning zone feedback specifically contradicting their answer - see Figure 9-34:

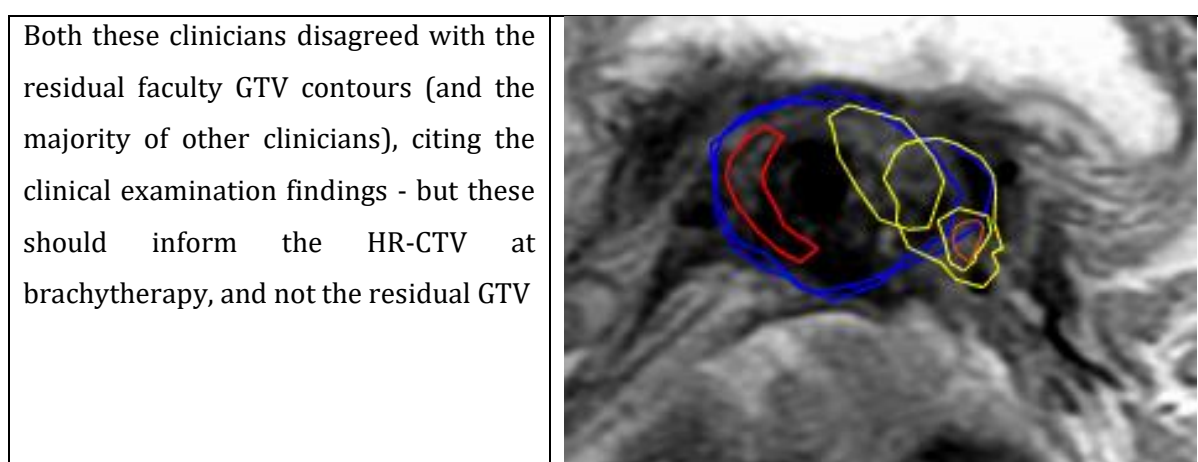


Figure 9-34 - Contours of two EMBRACE-II clinicians (blue) who disagreed with the 'exclude cervix' learning zone for the residual GTV

Other qualitative comments

23/52 (44%) post-workshop survey respondents submitted general comments and/or suggestions for improvements which were grouped into 33 themes by content analysis - quantitative results are displayed in Appendix Table A.9-7.

Most commonly (n=12) they included generally positive sentiment; 3 clinicians explicitly asked for more practice over time, for example: "it's good, keep on with interaction, surveys, testing and improving!", whereas one was luke-warm: "fine".

Four commented on the rapidity of practice - for example: "Easy compared to full contouring, more focus on principles in contouring than in individual slides" with one sounding a note of

caution: “has the danger of simply drawing without extra checks, manuals etc before you hit submit (in the home setting/ outside a course)”.

Two found the multiple reference contours confusing - for example: “I didn't understand the two yellow contours - were they 95% confidence intervals of “experts”? A single contour of the (average) expert's opinion might have been simpler, although less “true””.

One clinician suggested intermediate learning zone assessments as opposed to dichotomous pass/fail: “For some learning zones, it may be useful to have a 'yellow' warning instead of 'red'. For example, if one included the anterior vagina in CTV-IR (but maybe not the same extent as the expert), that could receive a yellow warning”.

9.6.3 Confidence & performance

Confidence

Clinicians' baseline confidence was higher for EBRT target volumes (mean 4.3/5) than for brachytherapy target volumes (3.7, $p < 0.01$, Mann-Whitney-U test), and lower for brachy target volumes than brachytherapy organs at risk (3.7 vs 4.4, $p < 0.01$). EBRT confidence was significantly higher than for the UK and international trainees for all regions of interest.

Paired changes in confidence for clinicians who completed both the pre- and post-workshop surveys ($n=43$) are shown in Figure 9-29. Average confidence increased very slightly in most tested volumes (mean increase = 0.16/5 points) but any increases were not statistically significant. Reported confidence *decreased* for some clinicians ($n=4$ (9%); range 2-7/43 = 4-19%) on each target volume.

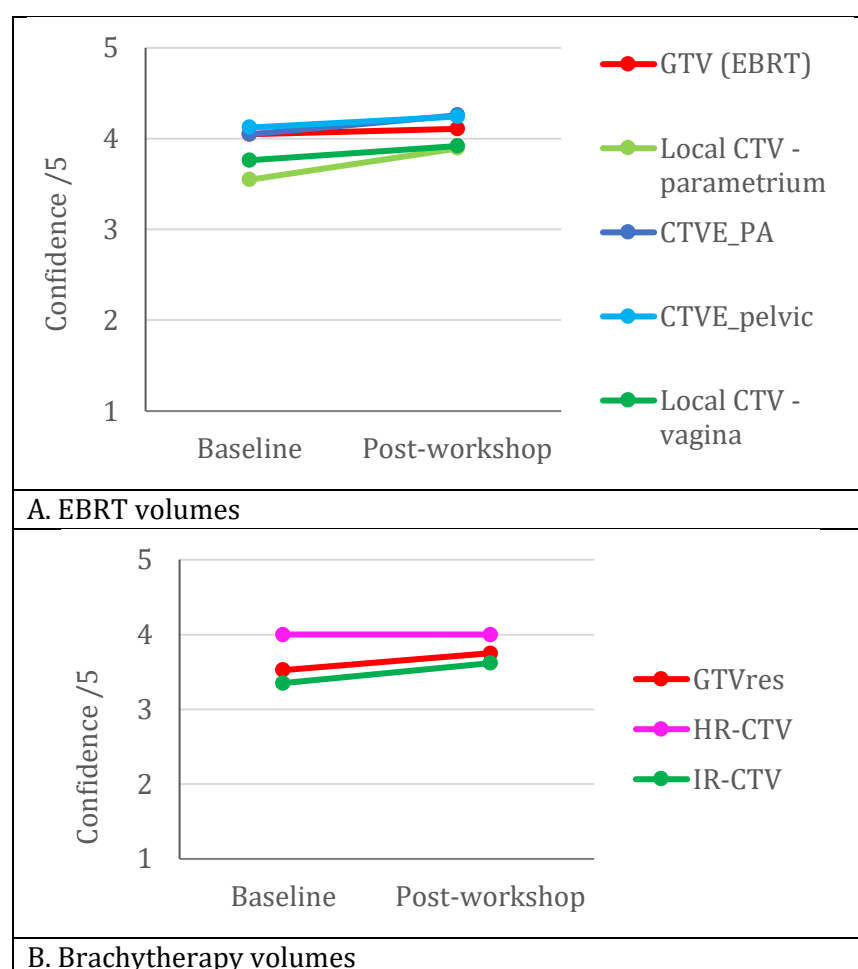


Figure 9-35 - EMBRACE-II: reported confidence (pre- and post-workshop) for target volumes tested

Confidence vs performance

Confidence was either not statistically significantly correlated with performance or weakly correlated (Appendix Table A.9-8). Most correlation coefficients were of negligible to weak strength and there were a mixture of positive and negative values.

Performance - comparison with trainee cohorts

Figure 9-36, Table 9-6 and Table 9-7 compare performance between the EMBRACE-II clinicians and trainee cohorts, assessed by JCI and mean learning zone score, for the three exercises in common.

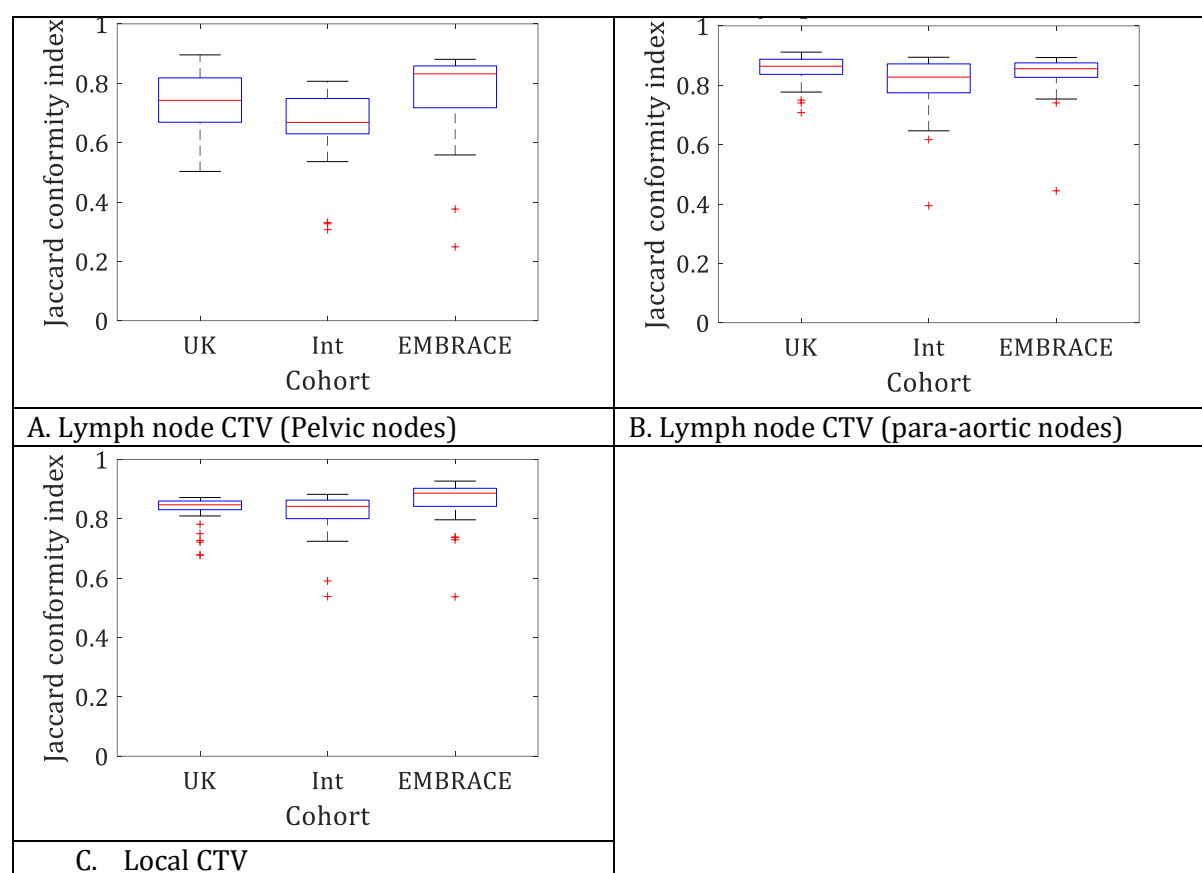


Figure 9-36 - Conformity for exercises repeated across all three cohorts. Int cohort = international trainees. For the UK trainees the para-aortic nodes exercise was an immediate repeat.

Table 9-6 - Conformity per cohort for the three repeated exercises

Region of interest	Median Jaccard conformity index			p
	UK trainees	International trainees	EMBRACE	
Elective lymph nodes - Pelvic	0.74	0.67	0.83	<0.01
Elective lymph nodes - Para-aortic	0.86*	0.83	0.86	<0.01
Local CTV	0.85	0.84	0.89	<0.01

Table 9-7 - Mean learning zone score for the three repeated exercises across the 3 pilot studies

Region of interest	Mean learning zone score			p
	UK trainees	International trainees	EMBRACE	
Elective lymph nodes - Pelvic	0.74	0.64	0.88	<0.01
Elective lymph nodes - Para-aortic	0.98*	0.84	0.93	<0.01
Local CTV	0.89	0.84	0.90	0.29

* these exercises were 'repeats' for the UK trainees

Conformity in the EMBRACE-II group was slightly higher than the trainee cohorts ($p < 0.01$ for all interactions), except for the UK trainees' para-aortic nodal conformity - this exercise was an immediate repetition for them.

Mean learning zone score was higher for EMBRACE-II in the pelvic nodes, but not in the para-aortic nodes or local CTV.

Performance - PIs vs non-PIs

There were generally no significant differences between the scores of PIs and non-PIs (Appendix Table A.9-9). There was a trend towards a lower score for PIs vs non-PIs in the brachytherapy residual GTV exercise - although this met statistical significance ($p = 0.016$) it may have been due to chance because of multiple testing.

There were insufficient participants to compare the performance of the PI at each centre with their colleagues.

Performance - replication of accreditation errors

The rates of learning zone errors in these exercises compared to the rates of the same types of error made in the accreditation cases are shown below in Table 9-8. Errors seen in 3-D simulation with the Addenbrooke's Contouring Tool were re-produced in Mini-Contour - there was a strong correlation between these respective rates (Spearman's $\rho = 0.78$, $p < 0.01$).

Table 9-8 Rates of learning zone errors in the EMBRACE-II workshop compared to rates of the same error in the accreditation exercises

Modality	Region of interest	Learning zone	Error in accreditation (% clinicians)	Error in learning zone (% clinicians)
EBRT	Elective lymph node CTV - pelvic nodes	exclude psoas muscle	0%	2%
		exclude psoas muscle	0%	0%
	Elective lymph node CTV - para-aortic nodes	exclude vertebral body	0%	0%
		include left lateral para-aortic lymph node region	35%	18%
	Local CTV	include para-vaginal Tissue	36%	58%
		exclude urethra	10%	6%
		exclude area lateral to parametrial border (in nodal volume)	8%	10%
		include all of parametrium	7%	8%
		exclude bladder	10%	0%
		exclude mesorectum	15%	23%
Brachy-therapy	GTVres	exclude normal cervix tissue	59%	37%
		exclude normal parametrium	8%	0%
	HR-CTV	include pathologic tissue ("grey zones")	22%	69%
		include whole cervix	22%	16%
		include distal parametrium if involved at diagnosis	27%	75%
	IR-CTV	exclude rectum	0%	0%
		include anterior vaginal wall (involved at diagnosis)	33%	52%
		exclude bladder wall	11%	38%
		exclude applicator and packing	30%	44%
		exclude extra-uterine tissue if uninvolved at diagnosis	31%	31%

9.7 Discussion

Exercise creation

Three of the faculty were able to agree on delineation errors and appropriate feedback, with only a minority of 'inclusion' and 'exclusion' learning zones requiring revision. The greater number of

edits required for exclusion zones adds further weight to the argument for their reconfiguration as discussed in Chapter 8. The degree of consensus is likely to vary between tumour sites, protocols/guidelines, cases, and target contouring versus organs at risk - this can be examined in future work.

Each exercise took approximately 1 hour to prepare and typically generated around 5-10 minutes of workshop activity or 3-5 minutes of self-directed learning i.e. very roughly 12 hours of faculty time per hour of user content. For Mini-Contour exercises this ratio has the potential to be reduced by automation (for example DICOM image upload and inbuilt case editing functionality) but a large (e.g. >50%) reduction seems unlikely. The resources required ("cost") to produce simulation-based medical education are infrequently reported, and the available data suggests trade-offs are unique to the intervention studied, although generally low-fidelity simulations report similar efficacy at lower cost (Zendejas et al., 2013). Whilst there are no formal data for the resources required to produce exercises in high-fidelity contouring tools, my own experience (from teaching courses using two high-fidelity simulations) suggests that content preparation for Mini-Contour is more intensive per hour of user activity than for high-fidelity simulation. For existing high-fidelity simulations uploading existing DICOM dataset & consensus adjustment of the reference contour(s) may then generate several hours of learner activity. However the development and maintenance of high-fidelity simulations is substantial - over £50,000 in the case of Addenbrooke's Contouring Tool compared with <£1000 for the initial version of Mini-Contour. Collection of resource use data for both high- and low-fidelity simulation is feasible and would inform any future comparisons of effectiveness.

User experience

Timing

This study has demonstrated that rapid deliberate practice is feasible in radiotherapy contouring education, with contouring taking less than 3 minutes per case as envisaged during development. The contouring speed at follow-up for the trainee cohorts suggests that users retained their ability to use the tool straightforwardly even after a time interval.

The quick loading times & low data transfer demonstrated are conducive to a smooth user experience and may be especially advantageous for users without broadband internet access or who pay for the amount of data they use.

The wide range of timings per exercise for all cohorts suggest that some users are rushed in a standard workshop format whereas others are waiting or progressing ahead of the facilitator.

Self-directed exercises may allow users to make the best use of their own time, although this must be balanced against the reduced engagement and exercise completion rates (see below), and the unknown effects (whether positive, negative or mixed) of group learning for Mini-Contour.

For the self-directed sub-groups within the international trainees and EMBRACE cohorts the time taken to review Mini-Contour feedback provides a preliminary look at users' engagement with feedback when unperturbed by direct observation (Chapter 8) or the live workshop setting. The increased review time of EMBRACE participants suggests more engagement with feedback - especially so given that EMBRACE clinicians' cognitive load is likely lower, enabling them to process feedback more quickly. Given the small numbers and lack of follow-up for the EMBRACE clinicians it is not possible to examine whether increased review time (a proxy for engagement with feedback) translated into improved performance at a later stage, but this question could be posed in future research. Looking forward, a more detailed analysis of how clinicians process the feedback could be enabled by either in-tool monitoring of their clicks (e.g. requiring them to acknowledge feedback by clicking it away) or unsupervised screen recording with widely available (but somewhat costly) software developed for usability purposes (for examples see <https://userbrain.net/> and <https://lookback.io/>).

Live workshops

Although live workshops with participants bringing their own laptops was feasible, the numbers of participants forgetting or not being able to bring a laptop (see North London UK and EMBRACE results) means that in some settings there will be learners present who will lose out on the active contouring experience. Whether the impact can be reduced by sharing computers is unclear; some participants felt the workshop environment had a positive impact but although learners generally favour group learning in the feedback, the effect on skills is much more mixed and may be negative (Cook et al., 2013).

Usability & reported usefulness

Users generally found it easy to learn to use the tool, as shown by the time to contour and the high system usability scale scores. Many of the findings support initial results from Chapter 8. The high frequency of comments about the drawing/editing functionality mean that this should be a high priority for the next development cycle. In addition, the potential to link with other content (such as anatomy, atlases and other contouring guidance) is important and relatively straightforward functionality that was highlighted by the UK trainees. In addition, issues were uncovered in these pilots that were not revealed by the detailed usability study, such as the contrast function not working for some users - this may be a browser compatibility issue and will be investigated.

Increasing age generally impedes user experience and usage of technology (Burton-Jones and Hubona, 2006) so it was interesting to see that the EMBRACE group (who, as experienced clinicians, we safely can infer are older on average) reported higher usability and similar usefulness compared to the two trainee cohorts. It is well-established that experts have reduced cognitive load compared to novices when navigating clinical problems (Sewell et al., 2019) so one may speculate this influenced cognitive bandwidth and made using Mini-Contour feel more straightforward for the EMBRACE cohort. Alternatively EMBRACE clinicians may be pre-selected as an especially enthusiastic and innovative group (they are at the forefront of innovation in cervix cancer radiotherapy) and so more likely to respond positively. Further testing in cohorts of experienced clinicians, including self-reported cognitive load measures, may help to clarify this and could initially be conducted at the GEC-ESTRO image-guided radiotherapy in gynaecological cancer course where I am a guest faculty member.

There was considerable overlap between suggestions for improvement between the three cohorts, with few new themes emerging from the EMBRACE group. This suggests that we may be approaching 'thematic saturation' (Saunders et al., 2018) and extensive further sampling to identify major areas for improvement would not be fruitful.

Fidelity

Fidelity or functional task alignment (percentage 'similarity to contouring in real life') was generally perceived to be high by all groups, although there was a significant range of responses. Especially surprising were responses that indicated 100% similarity - one suspects that these were expressions of enthusiasm and/or acceptability rather than critical judgements of the differences between Mini-Contour and contouring in clinical practice. The wording of the question (which could have used the phrase 'in your clinical practice') may have affected responses. Users' comments regarding fidelity were all acknowledged limitations of Mini-Contour design rather than new insights.

Also surprising was that EMBRACE clinicians' perceptions of fidelity were not significantly lower than trainees'. As with usability this may be a reflection of enthusiasm but is a promising indicator of acceptability in this group which contrast with previous users' (Chapter 8) expectations of senior clinicians using this software.

An acknowledged limitation (see Chapter 7) highlighted by participants is that, although quick, Mini-Contour exercises do not test the learner's ability to define the top and bottom slices of a particular volume, potentially a large source of variation in contouring (Eminowicz and McCormack, 2015). To enable rapid deliberate practice, alternatives to contouring a full 3-D case would be quiz questions where learners access a 2-D or 3-D image set and select the top and bottom slices of a given volume, or partial contouring (e.g. 3 selected slices: top, middle and bottom).

Engagement

Although the trainees were generally enthusiastic about the tool, the level of engagement with the self-directed (international trainees) and follow-up exercises (UK & international trainees) was moderate to low despite trainees being prompted via email. Possible factors include the timing of releasing the exercises and reminders, the programme's relevance to their current rotation (only a small minority of clinical trainees are in gynaecology rotations at any one time), and a relatively weak incentive for completion (the exercises were optional, with only a certificate for completion). Other studies have offered financial rewards for completion of follow-up exercises (Gillespie et al., 2017); whilst resources prohibit this routinely, a similar effect may be achieved if learners were to 'subscribe' to access the programme and then receive a partial refund on completion.

In the EMBRACE Group more non-PIs were engaged with Mini-Contour exercises than seen for high-fidelity during accreditation, but overall nearly half of eligible clinicians did not participate. Findings from Chapter 5 on the ability of clinicians to self-assess contouring competence suggest that engaging all clinicians in ongoing 'assessment for learning' is preferable to leaving them to decide whether they require further training.

Qualitative work exploring clinicians' engagement and underlying motivation (as performed in surgery, for example by Blackhall et al.(Blackhall et al., 2019)) is required to gain further insight.

For trainees, integrating a programme of contouring simulation into the relevant training attachment, as many users suggested, may increase engagement. This would require a co-ordinated effort amongst site-specialist clinicians, and the EMBRACE group and their trainees would be a promising cohort in which to pilot this.

As discussed in Chapter 8, strategies to promote the engagement of wider stakeholders such as professional bodies and training directors (both regional and local) in simulation is vital in trying to ensure maximum utilisation. Enabling 'transcripts' of trainees performance is one such

strategy, and could be printed out for trainees' portfolios (carrot) or sent directly to trainers if appropriate permissions were established in advance (stick).

Learning zones

Nearly all users reported that they generally agreed with the learning zone feedback and found it useful or very useful, which is encouraging for this novel concept. However, as mentioned previously in this thesis, learner enthusiasm for teaching does not translate into increased learning (Kirschner and van Merriënboer, 2013, Uttl et al., 2017).

As high frequency feedback can boost immediate performance but impair retention and transfer, especially for advanced learners (Hatala et al., 2014), prospective evaluation of the feedback type & frequency is needed - an area of interest in the wider simulation literature (Henriksen et al., 2018). The frequency and type of feedback would be relatively straightforward to manipulate within a randomised controlled trial design using different software versions, providing that learners are recruited into a programme of assessment and teaching.

The conflict between their use for assessment and their use for feedback seen in Chapter 8 was again apparent, and users' comments about learning zone stringency lend weight to the arguments presented in Chapter 8 to reconfigure them towards less stringent assessment.

Data showing the exquisite sensitivity of some learning zones' pass rates (especially the exclusion zones) to small areas of overlap support this, as illustrated by Figure 9-18 - there seems little value in penalising learners for what is largely acceptable variation, especially when a large anterior margin will be added to the local CTV to compensate for organ motion. The intention with that particular learning zone was to check whether the trainees understand the difference between the bladder wall muscle (grey on MRI - to which increased dose can cause toxicity (Manea et al., 2018)) and the urine (white on MRI - dose here will not cause toxicity) but this could have been accomplished with a different exercise asking trainees to contour the bladder, and then making the bladder wall a 'comment zone' in local CTV exercise.

One option for added functionality would be for learners to self-assess whether their contours have adequately excluded specific organs at risk. This may reduce learner's frustration as well as potentially increasing self-regulation (Panadero et al., 2017); indeed self-assessment is standard 'first-step' for debriefing after in-person healthcare simulation (Eppich and Cheng, 2015). The effect on learning of seeking to encourage meta-cognitive strategies is not universally positive however (Papinczak et al., 2008); as with feedback type and frequency above, adding a layer of

self-assessment to learning zones to enable a randomised-controlled comparison in a future study is feasible.

An even more ambitious goal would be to supplement learning zones with feedback about dosimetric effects. This requires sophisticated technology in the form of automated planning and its display to the learner but now looks to be possible, if not yet practically realised in the educational setting (Lim et al., 2019, Murphy and Gillespie, 2019).

The initial indication from the usability study that trainees were reluctant to comment on learning zones they perceived as incorrect was borne out in the trainee pilots, with a paucity of within-tool learning zone comments despite encouragement from the workshop facilitators. With accredited clinicians significantly more likely to comment and identify content errors or issues, trialling all content with groups of site-specialist clinicians is likely to provide the most robust post-production validation, although it is unlikely to be feasible for all scenarios. Shared rather than siloed institutional content can facilitate this (Caswell et al., 2008).

Performance measurement & lack of evidence for skill retention and transfer

As the pilots were designed as prospective cohort / case-control studies, in keeping with many educational design research evaluations (McKenney and Reeves, 2020), there are many factors which may have confounded measurement of participants' contouring performance. These include: variations in exercise difficulty and the placement of learning zones across different cases, the effects (sub-conscious or otherwise) of me facilitating the workshop as opposed to feedback from the tool, and the effects of a prior lecture on anatomy and radiology for two out of three UK trainee groups. In future work these variables could be partially addressed or mitigated by:

- Studying self-directed learning with simulation practice embedded in a wider education programme
- "Counterbalancing" i.e. systematically varying the order of exercises within themes to produce objective data regarding exercise difficulty
- Using learning zones that have been extensively validated

Re-testing the same exercise at follow-up which is common in contouring research (see Chapter 2), whilst eliminating some of these factors, may be a measurement of 'recall' whereas meaningful learning is demonstrated by the ability to transfer principles to a different situation after a time interval (Shariff et al., 2020).

Congruence of trends in performance by JCI & learning zones seen in the UK trainee pilot is reassuring in that they are measuring skill as an underlying construct. The sensitivity of the pass rate to learning zone stringency, especially for exclude zones, could be a result of measurement error i.e. placement of a learning zone in an area of acceptable variation. A way to test this hypothesis would be for experts to score the contours manually without reference to the learning zone and detect the point at which learning zone inclusion/exclusion predicted expert score (if at all). If an acceptable overlap of minimum 90% for 'include' learning zones and maximum 2% for 'exclude' learning zones were to be used, these percentages would equate to a similar absolute area^{xii} of acceptable leniency, as the average size of include learning zones was five times smaller than that of exclude learning zones.

Confidence and performance

Although objective performance measurement had limitations, the lack of relationship between clinicians' confidence and their performance seen in this study was foreseeable. This phenomenon was described by Dunning and Kruger in 1999 (Kruger and Dunning, 1999): their seminal paper describes the weak relationship between participants' perceptions of their humour and logical reasoning (amongst other domains) and the results of objective tests. This effect is also widespread in medicine - a JAMA meta-analysis (Davis et al., 2006) of studies reporting comparisons of self-assessed versus observed competence reported:

"Of the 20 comparisons between self- and external assessment, 13 demonstrated little, no, or an inverse relationship and 7 demonstrated positive associations. A number of studies found the worst accuracy in self-assessment among physicians who were the least skilled and those who were the most confident."

An increase of confidence with stage of training may reflect UK trainees' familiarity with the subject material and/or their expectations of themselves in general. The weak increase of performance with increasing clinical experience was more surprising, and may have been reduced by the pre-workshop lecture in the UK cohort, but the finding held true for international trainees. The implication for a programme of radiotherapy contouring competency is that baseline confidence (and perhaps also experience) should not be used as a basis for selecting the level of difficulty at which participants enter.

^{xii} This area is measured in screen pixels rather than related to patient characteristics as the two were not linked in Mini-Contour v1.0.

EMBRACE clinicians appeared to correctly sense that a one-off contouring workshop was unlikely to significantly improve their contouring skill as their self-efficacy stayed stable. This may be linked to more advanced meta-cognitive abilities but responses may also be biased by ‘anchor bias’ (i.e. survey responses pulled closer to the middle item in Likert-like scales) in the questionnaire. The increase in confidence for target volumes that were not taught or tested results from the “halo effect”, another well-known cognitive bias prevalent in education (Nisbett and Wilson, 1977, Boet et al., 2012) where participants evaluation of one aspect of an educational evaluation can colour their perception of a separate aspect.

Performance over time

Initial performance improvements seen in the UK workshops for inclusion learning zones were encouraging, especially as these relate to ‘hitting the target’ by ensuring appropriate coverage, whereas exclusion zones relate to sparing normal tissues by reducing dose to the organs at risk. Although the lack of retention of this improvement at 4 weeks may seem discouraging, decay of knowledge and skills after initial acquisition is a well-studied phenomenon, dating back to Ebbinghaus’ “forgetting curve” in 1885 (Ebbinghaus, 1885). Skill decay has been reproduced in a wide variety of psychological domains (Arthur et al., 1998), including medicine (Sinha et al., 2008, Lammers, 2008).

The UK pilot highlights the importance of testing skill retention where initial improvements have been seen, especially as at follow-up confidence remained erroneously elevated: increased confidence with no change in performance is arguably the worst possible outcome for a teaching intervention! These findings also challenge the setup of many existing post-graduate teaching programmes, which often only test knowledge or skill acquisition on a single training day, if they measure performance improvement at all.

Repetitive practice at increasing intervals - “spaced practice” - has been shown to promote long term skill retention better than practising all at once - “massed practice” (Larsen et al., 2008, Cecilio-Fernandes et al., 2018). We can conclude from the longitudinal programme (even in the presence of some considerable measurement uncertainty) that spaced repetition of 5 short exercises in unselected international trainees is *insufficient to ensure complete retention and transfer* of contouring principles at delayed testing.

For the repeated exclusion learning zones, trainees’ consistent high performance on at least two out of five suggest that they could be removed from the exercise - giving positive feedback may provide false reassurance regarding general competence, as hinted at in Chapter 8 (Figure 8-11).

Replication of EMBRACE accreditation errors

The EMBRACE group pilot demonstrated that errors seen in high-fidelity simulation can be replicated in low-fidelity, as hinted at in early user trials (Chapter 7). This supports the assertion that Mini-Contour could be used as quick strategy to highlight and remediate common errors in contouring, both within and outside clinical trials, but evidence that this produces a lasting change in practice is needed. Further longitudinal data from education and testing within EMBRACE-II may provide some indication of this.

Mini-Contour: ready for a definitive trial?

Given the naturalistic nature of this observational pilot study, drawing solid inferences about effects of Mini-Contour on learning is unadvisable. However, before a controlled trial to assess efficacy is performed significant questions remain, including: 'how can we optimise clinicians' engagement in a contouring learning programme?', 'what are clinicians reasoning processes and how do they affect contouring variation', 'what is the optimum way of providing feedback and does this differ between novice and advanced learners?'. Clarification is needed (Cook et al., 2008) by studies addressing these questions before we proceed to a definitive trial. Many questions can benefit both high and low-fidelity simulation programmes; possible future studies are outlined in Chapter 10.

Implications of these findings for future development

The immediate next step in development after further user experience testing in experienced clinicians will be to meet with study investigators and the development team to collate usability issues from both the pilot and usability studies, evaluate the resources required for the desired improvements against those available, and decide on immediate priorities. These are likely to include:

- Technical development of the simulation including an enhanced drawing and editing interface, and linkage to other resources
- Development of a way to test and teach the identification of top and bottom slices of regions of interest
- A reconfiguration of learning zone display, assessment and stringency
- Situating the simulation(s) within learning exercises, including pre-learning for junior trainees
- Examining the effects of spaced practice within a relevant tumour site rotation
- Formative evaluation in other tumour sites (for example head and neck cancer)

These developments *and* further research should be enacted prior to a definitive trial to compare the effectiveness of Mini-Contour and high-fidelity simulation (the current standard of care) for contouring skill improvement.

9.8 Conclusion

These pilot studies have demonstrated the feasibility and acceptability, to both trainees and accredited practitioners, of using a low-fidelity simulation to teach radiotherapy contouring via a deliberate practice approach. It has shown strong clinician support (both trainees and accredited practitioners) for the principle of 'learning zones' as automated feedback.

The tool enabled assessment of contouring across a broad range of exercises in a limited time and could form the core of a radiotherapy contouring deliberate practice programme if paired with effective instructional techniques. The findings highlight that practicing skills and measuring skill retention and transfer over time is a vital part of any teaching intervention - a challenge to the conduct of much of post-graduate training - and that the simulation must be seen within an overall programme and curriculum that is relevant to the learner's goals.

Further research and development is needed before a definitive trial comparing Mini-Contour to high-fidelity simulation.

10 Conclusion

This final chapter first draws together the ways in which this programme of educational design research has addressed my overarching research questions and made original contributions to the literature. It then outlines possible avenues for future development and research.

Box 10-1 - A restatement of the overarching research questions for this programme of research

- How can medical education literature and the wider educational literature regarding simulated practical skills training inform our approach to the teaching and assessment of radiotherapy contouring?
- How can this knowledge be applied to shape the simulated assessment and teaching of radiotherapy contouring?
- What are the impacts of novel approaches using web-based technology on the teaching and assessment of radiotherapy contouring in the 'real world'?

Chapters 2-3 and 5-6 form the 'analysis and exploration' phase of this educational design research project. Chapter 2 made the case that, over 20 years after the introduction of advanced radiotherapy techniques, contouring variation remains a potential weak link in the delivery of high-quality radiotherapy. Several strategies have been utilised to minimise contouring variability, of which the most consistently successful has been detailed contouring protocols illustrated with worked examples (radiotherapy 'atlases'). Multi-modal imaging, automated segmentation, peer review and training programmes have also been utilised but sometimes with mixed or uncertain impact. For educational interventions there is little systematic research into whether any learning is retained, and which underlying constructs they are attempting to engage with and/or modify.

Chapter 2 also highlighted some of the weaknesses in radiotherapy quality assurance assessment processes, which is the main arena for summative assessment of radiotherapy contouring. Due to resource limitations, contouring quality assurance is often performed on one or two cases, which may not be representative of the spectrum of cases in a trial (i.e. limited content validity). Generally only principal investigators are assessed, and there is little standardisation of assessment processes or information about their reproducibility. If clinicians fail the assessment then they can re-submit the same case.

Chapter 3 explains how issues such as those with assessment validity have been addressed in the educational literature and how this body of knowledge is helpful in addressing some of the weaknesses in contouring assessment (see section 3.7.3). It also revealed the relevance of other areas established educational theory and best practice to radiotherapy contouring education, much of which has not been consciously applied to this context. The simulation literature suggests that low-fidelity simulation may work as well or better than high-fidelity simulation for novices to improve practical skills, and that the resources and software complexity required can be significantly lower. Cognitive load theory provides a possible explanation for these effects, as well as for the improvement of contouring seen with radiotherapy atlases. Further application of cognitive load theory principles to radiotherapy contouring education (Table 3-3) has the potential to improve the effectiveness of contouring interventions. Deliberate practice theory provides a framework for explaining how clinicians can progress from contouring novice to expert within a structured process of simulated assessment and feedback, although much work remains before such a programme can be realised (section 3.6.4).

Given the necessary limitations of the scope and depth of the review of educational literature, there are potentially many more promising applications for radiotherapy education within and outside of contouring. This supports increased training of radiotherapy educators in educational theory, best practices and research methods - especially regarding skills training, assessment and feedback.

The case studies in radiotherapy contouring quality assurance for the EMBRACE-II trial (Chapters 5 & 6) illustrate some of the limitations in current contouring quality assurance practices and the possibilities to learn lessons from educational best practice (section 6.4). They also show that high-fidelity assessment and teaching, despite a high level of realism, may bring constraints regarding the time taken and coverage of the breadth of clinical variation seen in clinical trials and routine practice.

Common errors in external beam radiotherapy and image-guided adaptive brachytherapy contouring for cervix cancer were identified. These form the basis of targeted learning exercises later in the thesis, where it was shown that these errors are repeated outside of the quality assurance process (Chapters 7 & 9).

Chapter 5 also evaluated a novel online continuing education programme integrated into the EMBRACE-II trial radiotherapy quality assurance. Even experienced clinicians were not always able to predict their own learning needs, which emphasises the importance of formative

assessment of radiotherapy contouring. Further work is required to explore the factors behind the low to moderate engagement with self-directed learning materials.

In Chapters 5 and 6 data were also presented to critique the use of conformity indices in individual assessments as a surrogate for expert appraisal. Conformity index is insufficiently discriminatory when attempting to distinguish between expert-assessed adequate and inadequate contours. The findings presented also challenge the prevailing use of standard conformity index cut-off across different regions of interest and cases. Further work could explore the relationship between other metrics (such as surface distance) as this analysis was limited to geometric overlap. An ideal automated assessment would be able to effectively distinguish between competent or non-competent clinicians whilst at the same time providing useful feedback.

Chapter 7 marked the end of 'analysis and exploration' and the beginning of the 'design and construction' phase where the findings above were incorporated to shape the design of a new intervention. Chapters 2, 3, 5 and 6 showed that there is need for learners to be able to practice applying contouring concepts across cases, targeting errors, in a time-efficient manner - this need can be addressed by low fidelity simulation. The development process has shown that low-fidelity contouring simulation of radiotherapy contouring with automated feedback is feasible, and can enable rapid cycling through contouring exercises, breaking down contouring a whole case into small tasks (Chapters 7-9). This can enable a greater understanding of where clinicians are making errors, and also increase assessment validity through contouring over a wider number of clinical scenarios. Targeted qualitative feedback is a powerful facilitator of learning and a requirement for deliberate practice (see sections 3.6 and 3.8) - in Chapter 7 the concept of 'learning zones' was described to enable this at scale, given the resource burden of manual assessment of large numbers of learners.

Low-fidelity simulation allowed pedagogical innovation with flexibility of the code base and low development costs. Mini-Contour was produced on a software development budget of less than £1000, which compares favourably with >£50,000 required to develop the high-fidelity Addenbrooke's Contouring Tool. The low amount of data transferred for each exercise may have benefits for learners in countries where the internet is slow or access costly.

Chapters 8 and 9 mark the first 'evaluation and reflection' phase of this educational design research programme. In the qualitative study in Chapter 8 users adapted to the software quickly and reacted positively to the novel elements of rapid practice and learning zone feedback. Their detailed comments suggested how learning zones could be refined conceptually and

technologically. This study also highlighted clinical reasoning and meta-cognitive processes (e.g. self-regulation), that should be explored in more depth in future research.

The pilot studies in Chapter 9 showed the ability of low-fidelity contouring simulation to rapidly test specific contouring concepts and reproduce the errors seen in high-fidelity simulation - an important proof of principle. Groups of trainees (both UK and international) and accredited clinicians perceived Mini-Contour to be easily usable and highly useful. Surprisingly, all pilot groups felt the simulation was highly similar to real life - this suggests good functional task alignment despite the simplicity of the software. The lack of relationship between contouring confidence and performance highlights the need for formative assessments to attempt to realign these. In the pilot studies the learning zone feedback was generally received enthusiastically by trainee and accredited clinicians alike but their validity for assessment and utility for learning remains to be demonstrated as part of future work.

Analysis of retention and transfer of contouring skill is rarely reported in contouring education (Cacicedo et al., 2019) but was included in the trainee pilot studies in Chapter 9. The lack of contouring skill retention and transfer and moderate trainee engagement over time seen in the trainee pilot studies is a challenge to current 'training day' model in radiation oncology and a reason to consider a longitudinal approach within clinical training rotations for future interventions and research.

Usability data from both studies identified multiple issues that could be addressed by further development - both technical and pedagogical. Improvements to the contouring interface, links to other resources, automated case creation & editing, and testing top & bottom slices are high on the list of priorities. On the pedagogical front more work is required including establishing exercise difficulty as part of crafting a contouring curriculum, and learning resources to frame the simulation exercises.

A key characteristic of educational design research is that the results of the evaluation phase inform existing educational theory (Chapter 4). With regard to the medical educational literature, this requirement has not been met - during this programme of research I have focussed on applying educational theory and best practices to the radiotherapy contouring domain. To the extent that theoretical contributions have been made, they are 'humble' (Bakker and van Eerde, 2014, DiSessa and Cobb, 2004) i.e. they attempt to advance theory and practice in the specific domain of radiotherapy contouring education. As Bakker and van Eerde put it: *"it is very rare that a theoretical contribution to aerodynamics will be made in the design of an airplane; yet innovations in airplane design occur regularly"* (Bakker and van Eerde, 2014, p.13-14). In future work as the

intervention matures, it may well be possible to ask questions (for example about feedback, practice, self-regulation, motivation, performance and learning in post-graduate trainees and certified practitioners) whose answers would contribute to the wider body of medical education research.

Future studies

The end goal of this programme of research would be to test these innovations against the current standard of care in contouring education (high-fidelity simulation) as part of a randomised-controlled trial with skill (i.e. simulation score) or behavioural (i.e. clinical contour score) outcomes as the primary endpoint (see Table 3-1). It might even be possible to measure the impact on clinical outcomes, the standard to which other cancer interventions are held, of a multi-faceted intervention given the ability of clinical trial groups such as the EMBRACE group to register and analyse clinical outcomes relatively efficiently (Tan et al., 2020).

Before that is planned, however, more exploratory and explanatory research is needed. Many interventions in medical education have floundered due to a lack of understanding ‘how and why’ they might work (Cook et al., 2008). Four possible studies are outlined briefly below.

Investigation of the clinical reasoning mechanisms underlying contouring decisions is an important next step - *“the foundational nature of clinical reasoning across professions makes research ..., teaching ..., and assessment ... of clinical reasoning essential”* (Young et al., 2018). This is an area of active research in the health sciences generally that has not been applied to radiotherapy. A qualitative study could be designed to explore clinical reasoning processes in simulated and naturalistic settings with both experienced and novice clinicians - this study would expand upon the think-aloud methodology seen in Chapter 8, incorporating script theory and situativity theory in its theoretical framework.

A theory-based exploration of clinicians’ motivation for engaging (or not) with contouring education programmes is also important, as limited engagement with available learning materials was a theme in Chapters 5 & 9. Such mixed-methods studies have been conducted for surgical simulation training (Blackhall et al., 2019) and should be repeated in the radiotherapy contouring domain.

Once the next ‘design and construction’ phase is complete, an obvious next ‘evaluation and reflection’ study would be to create a longitudinal contouring training programme within a specific training rotation to examine learners’ engagement and skill changes during this period. Assessment of the impact of the tool would be confounded by other training and experience, but

would allow evaluation of the contouring programme in a setting where participants' intrinsic motivation is likely to be higher than on isolated training days.

An exploration of the varying effects of qualitative and quantitative feedback in different groups (i.e. experts and novices) would provide valuable insights about the effect of such feedback on learning in different learner groups and contexts. This would be suited to a randomised study with manipulation of the type and frequency of feedback per exercise between groups, and could be situated within a longitudinal programme described above if a sufficient number of learners can be recruited. The effect of this type of feedback on different learner groups is poorly understood even in the wider medical education field (Henriksen et al., 2018) and so findings would have wider relevance than just for radiotherapy contouring.

The most obvious barrier to these plans is a lack of resources. Other hurdles include stakeholder engagement, and scalability. Collaboration with other interested research groups (Murphy and Gillespie, 2019, Evans et al., 2019b) to overcome these could be fruitful. Further funding is needed, not only for technological development but also (arguably more importantly) to provide dedicated research time (Ajjawi et al., 2018). Rigorous medical education research with meaningful clinical endpoints is unlikely to be completed without a significant investment in durable programmes of research conducted by experienced and dedicated researchers (McGaghie et al., 2014). Recent efforts to develop such capacity in the UK are encouraging (National Institute for Health Research (UK), 2020).

References

- Abrams, R. A., Winter, K. A., Regine, W. F., Safran, H., Hoffman, J. P., Lustig, R., . . . Willett, C. G. 2012. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704--a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. *International Journal of Radiation Oncology, Biology, Physics*, 82, 809-16.
- Ackerman, P. L. 2014. Nonsense, common sense, and science of expert performance: Talent and individual differences. *Intelligence*, 45, 6-17.
- Agnew, A. J. & O'kane, C. J. 2011. Addressing the hiatus of learning incentives for prevocational doctors: continuing medical education points for interns. *Medical Journal of Australia*, 194, 349-352.
- Ahmad, S. S., Duke, S., Jena, R., Williams, M. V. & Burnet, N. G. 2012. Advances in radiotherapy. *BMJ (Clinical research ed.)*, 345, e7765.
- Ajjawi, R., Crampton, P. E. S. & Rees, C. E. 2018. What really matters for successful research environments? A realist synthesis. *Medical Education*.
- Alfieri, L., Brooks, P. J., Aldrich, N. J. & Tenenbaum, H. R. 2011. Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103, 1-18.
- American Educational Research Association 2014. *Standards for educational and psychological testing*, American Educational Research Association American Psychological Association ...
- Anastakis, D. J., Regehr, G., Reznick, R. K., Cusimano, M., Murnaghan, J., Brown, M. & Hutchison, C. 1999. Assessment of technical skills transfer from the bench training model to the human model. *The American Journal of Surgery*, 177, 167-170.
- Anderson, T. & Shattuck, J. 2012. Design-Based Research: A Decade of Progress in Education Research? *Educational Researcher*, 41, 16-25.
- Arksey, H. & O'malley, L. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8, 19-32.
- Arthur, W., Bennett, W., Stanush, P. L. & Mcnelly, T. L. 1998. Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance*, 11, 57-101.
- Atkinson, R. C. & Shiffrin, R. M. 1968. Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, 2, 89-195.
- Bakker, A. 2019. *Design research in education : a practical guide for early career researchers*, New York, Routledge.
- Bakker, A. & Van Eerde, D. 2014. An introduction to designbased research with an example from statistics education. In: Bikner-Ahsbahs, A., Knipping, C. & Presmeg, N.(eds.) *Doing qualitative research: methodology and methods in mathematics education*. New York: Springer.
- Bangor, A., Kortum, P. T. & Miller, J. T. 2008. An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24, 574-594.

- Barab, S. 2014. Design-Based Research: A Methodological Toolkit for Engineering Change. In: Sawyer, R. K. (ed.) *The Cambridge Handbook of the Learning Sciences*. 2 ed. Cambridge: Cambridge University Press.
- Barab, S. & Squire, K. 2004. Design-Based Research: Putting a Stake in the Ground. *Journal of the Learning Sciences*, 13, 1-14.
- Barnum, C. M. 2008. *What do you mean when you say "usability"?* [Online]. Available: <https://elearnmag.acm.org/archive.cfm?aid=1361077> [Accessed 10/12/2020].
- Barnum, C. M. 2011. *Usability Testing Essentials*, Elsevier.
- Barrows, H. S. 1968. Simulated patients in medical teaching. *Canadian Medical Association Journal*, 98, 674-676.
- Barsuk, J. H., Cohen, E. R., Feinglass, J., Mcgaghie, W. C. & Wayne, D. B. 2009. Use of simulation-based education to reduce catheter-related bloodstream infections. *Archives of Internal Medicine*, 169, 1420-3.
- Barton, M. B., Jacob, S., Shafiq, J., Wong, K., Thompson, S. R., Hanna, T. P. & Delaney, G. P. 2014. Estimating the demand for radiotherapy from the evidence: a review of changes from 2003 to 2012. *Radiotherapy and Oncology*, 112, 140-4.
- Beadle, B. M., Jhingran, A., Yom, S. S., Ramirez, P. T. & Eifel, P. J. 2010. Patterns of regional recurrence after definitive radiotherapy for cervical cancer. *International Journal of Radiation Oncology, Biology, Physics*, 76, 1396-403.
- Bekelman, J. E., Deye, J. A., Vikram, B., Bentzen, S. M., Bruner, D., Curran, W. J., Jr., . . . Purdy, J. 2012. Redesigning radiotherapy quality assurance: opportunities to develop an efficient, evidence-based system to support clinical trials--report of the National Cancer Institute Work Group on Radiotherapy Quality Assurance. *International Journal of Radiation Oncology, Biology, Physics*, 83, 782-90.
- Bell, L., Holloway, L., Bruheim, K., Petric, P., Kirisits, C., Tanderup, K., . . . Hellebust, T. P. 2020. Dose planning variations related to delineation variations in MRI-guided brachytherapy for locally advanced cervical cancer. *Brachytherapy*, 19, 146-153.
- Ben-David, M. F. 2000. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, 22, 120-130.
- Berman, N. B., Durning, S. J., Fischer, M. R., Huwendiek, S. & Triola, M. M. 2016. The Role for Virtual Patients in the Future of Medical Education. *Academic Medicine*, 91, 1217-22.
- Bing-You, R., Hayes, V., Varaklis, K., Trowbridge, R., Kemp, H. & Mckelvy, D. 2017. Feedback for Learners in Medical Education: What is Known? A Scoping Review. *Academic Medicine*, 92, 1346-1354.
- Black, P. & Wiliam, D. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5-31.
- Blackhall, V. I., Cleland, J., Wilson, P., Moug, S. J. & Walker, K. G. 2019. Barriers and facilitators to deliberate practice using take-home laparoscopic simulators. *Surgical Endoscopy*, 33, 2951-2959.

- Boell, S. K. & Cecez-Kecmanovic, D. 2014. A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches. *Communications of the Association for Information Systems*, 34.
- Boet, S., Sharma, S., Goldman, J. & Reeves, S. 2012. Review article: Medical education research: an overview of methods. *Canadian Journal of Anesthesia-Journal Canadien D Anesthesie*, 59, 159-170.
- Bradley, P. 2006. The history of simulation in medical education and possible future directions. *Medical Education*, 40, 254-62.
- Braun, V. & Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77-101.
- Brooke, J. 1996. SUS: a "quick and dirty" usability scale. In: Jordan, P., Thomas, B. & Weerdmeester, B. (eds.) *Usability evaluation in industry*. London, UK: Taylor & Francis.
- Brown, A. L. 1992. Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *Journal of the Learning Sciences*, 2, 141-178.
- Brydges, R., Carnahan, H., Rose, D., Rose, L. & Dubrowski, A. 2010. Coordinating Progressive Levels of Simulation Fidelity to Maximize Educational Benefit. *Academic Medicine*, 85, 806-812.
- Brydges, R., Hatala, R., Zendejas, B., Erwin, P. J. & Cook, D. A. 2015. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Academic Medicine*, 90, 246-56.
- Burnet, N. G., Thomas, S. J., Burton, K. E. & Jefferies, S. J. 2004. Defining the tumour and target volumes for radiotherapy. *Cancer Imaging*, 4, 153-61.
- Burton-Jones, A. & Hubona, G. S. 2006. The mediation of external variables in the technology acceptance model. *Information & Management*, 43, 706-717.
- Cacicedo, J., Navarro-Martin, A., Gonzalez-Larragan, S., De Bari, B., Salem, A. & Dahele, M. 2019. Systematic review of educational interventions to improve contouring in radiotherapy. *Radiotherapy and Oncology*, 144, 86-92.
- Caldwell, C. B., Mah, K., Ung, Y. C., Danjoux, C. E., Balogh, J. M., Ganguli, S. N. & Ehrlich, L. E. 2001. Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: The impact of 18FDG-hybrid PET fusion. *International Journal of Radiation Oncology Biology Physics*, 51, 923-931.
- Cancer Research Uk. 2016. *About radiotherapy* [Online]. Available: <https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/radiotherapy/about> [Accessed 19/10/2020].
- Cantillon, P. & Sargeant, J. 2008. Giving feedback in clinical settings. *British Medical Journal*, 337.
- Carty, E. 2010. Educating Midwives with the World's First Simulator: Madame du Coudray's Eighteenth Century Mannequin. *Canadian Journal of Midwifery Research and Practice*, 9.
- Caswell, T., Henson, S., Jensen, M. & Wiley, D. 2008. Open content and open educational resources: Enabling universal education. *The International Review of Research in Open and Distributed Learning*, 9.

- Cattaneo, G. M., Reni, M., Rizzo, G., Castellone, P., Ceresoli, G. L., Cozzarini, C., . . . Calandrino, R. 2005. Target delineation in post-operative radiotherapy of brain gliomas: Interobserver variability and impact of image registration of MR(pre-operative) images on treatment planning CT scans. *Radiotherapy and Oncology*, 75, 217-223.
- Cecilio-Fernandes, D., Cnossen, F., Jaarsma, D. & Tio, R. A. 2018. Avoiding Surgical Skill Decay: A Systematic Review on the Spacing of Training Sessions. *Journal of Surgical Education*, 75, 471-480.
- Chan, A. J., Islam, M. K., Rosewall, T., Jaffray, D. A., Easty, A. C. & Cafazzo, J. A. 2010. The use of human factors methods to identify and mitigate safety issues in radiation therapy. *Radiotherapy and Oncology*, 97, 596-600.
- Chandra, D. B., Savoldelli, G. L., Joo, H. S., Weiss, I. D. & Naik, V. N. 2008. Fiberoptic oral intubation: the effect of model fidelity on training for transfer to patient care. *Anesthesiology*, 109, 1007-13.
- Chen, A. M., Chin, R., Beron, P., Yoshizaki, T., Mikaeilian, A. G. & Cao, M. 2017. Inadequate target volume delineation and local – regional recurrence after intensity-modulated radiotherapy for human papillomavirus- positive oropharynx cancer. *Radiotherapy and Oncology*, 123, 412-418.
- Chen, A. M., Farwell, D. G., Luu, Q., Chen, L. M., Vijayakumar, S. & Purdy, J. A. 2011. Marginal misses after postoperative intensity-modulated radiotherapy for head and neck cancer. *International Journal of Radiation Oncology Biology Physics*, 80, 1423-1429.
- Chen, W. & Reeves, T. C. 2020. Twelve tips for conducting educational design research in medical education. *Medical Teacher*, 42, 980-986.
- Choi, J., Yoon, H. I., Lee, J., Keum, K. C., Kim, G. E. & Kim, Y. B. 2015. Optimal Extent of Prophylactic Irradiation of Paraaortic Lymph Nodes in Patients with Uterine Cervical Cancer. *PloS One*, 10, 11.
- Choudhry, N. K., Fletcher, R. H. & Soumerai, S. B. 2005. Systematic review: the relationship between clinical experience and quality of health care. *Annals of Internal Medicine*, 142, 260-73.
- Ciardo, D., Argenone, A., Boboc, G. I., Cucciarelli, F., De Rose, F., De Santis, M. C., . . . Leonardi, M. C. 2017. Variability in axillary lymph node delineation for breast cancer radiotherapy in presence of guidelines on a multi-institutional platform. *Acta Oncologica*, 56, 1081-1088.
- Cibula, D., Potter, R., Planchamp, F., Avall-Lundqvist, E., Fischerova, D., Haie-Meder, C., . . . Raspollini, M. R. 2018. The European Society of Gynaecological Oncology/European Society for Radiotherapy and Oncology/European Society of Pathology Guidelines for the Management of Patients with Cervical Cancer. *Virchows Archiv*, 472, 919-936.
- Cleland, J. 2015. Exploring versus measuring: considering the fundamental differences between qualitative and quantitative research. In: Cleland, J. & Durning, S.(eds.) *Researching Medical Education*. Wiley Blackwell.
- Cobb, P., Confrey, J., Disessa, A., Lehrer, R. & Schauble, L. 2003. Design Experiments in Educational Research. *Educational Researcher*, 32, 9-13.
- Cohen, L., Manion, L. & Morrison, K. 2017. *Research methods in education*, routledge.

- Cole, R., Purao, S., Rossi, M. & Sein, M. 2005. Being proactive: where action research meets design research. *ICIS 2005 proceedings*, 27.
- Collins, A. Toward a Design Science of Education. 1992 Berlin, Heidelberg. Springer Berlin Heidelberg, 15-22.
- Collins, A., Joseph, D. & Bielaczyc, K. 2004. Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, 13, 15-42.
- Conibear, J. R. 2018. *Assessment of Target Volume and Organ at Risk Contouring Variability within the Context of UK Head and Neck and Lung Cancer Radiotherapy Clinical Trials*. M.D. (Res) Thesis (Doctoral), University College London.
- Cook, D. 2014. How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. *Medical Education*, 48, 750-760.
- Cook, D., Hamstra, S., Brydges, R., Zendejas, B., Szostek, J., Wang, A., . . . Hatala, R. 2013. Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher*, 35, E844-E875.
- Cook, D., Hatala, R., Brydges, R., Zendejas, B., Szostek, J., Wang, A., . . . Hamstra, S. 2011. Technology-Enhanced Simulation for Health Professions Education A Systematic Review and Meta-analysis. *Jama-Journal of the American Medical Association*, 306, 978-988.
- Cook, D. A., Blachman, M. J., Price, D. W., West, C. P., Thomas, B. L. B., Berger, R. A. & Wittich, C. M. 2018. Educational Technologies for Physician Continuous Professional Development: A National Survey. *Academic Medicine*, 93, 104-112.
- Cook, D. A., Bordage, G. & Schmidt, H. G. 2008. Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Medical Education*, 42, 128-33.
- Cook, D. A., Brydges, R., Hamstra, S. J., Zendejas, B., Szostek, J. H., Wang, A. T., . . . Hatala, R. 2012. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. *Simul Healthc*, 7, 308-20.
- Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J. & Montori, V. M. 2010. Instructional Design Variations in Internet-Based Learning for Health Professions Education: A Systematic Review and Meta-Analysis. *Academic Medicine*, 85, 909-922.
- Cox, S., Cleves, A., Clementel, E., Miles, E., Staffurth, J. & Gwynne, S. 2019. Impact of deviations in target volume delineation - Time for a new RTQA approach? *Radiotherapy and Oncology*, 137, 1-8.
- Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. & Petticrew, M. 2008. Developing and evaluating complex interventions: new guidance.
- Cubillos Mesias, M., Boda-Heggemann, J., Thoelking, J., Lohr, F., Wenz, F. & Wertz, H. 2016. Quantification and Assessment of Interfraction Setup Errors Based on Cone Beam CT and Determination of Safety Margins for Radiotherapy. *PLoS One*, 11, e0150326.
- Daniels, V. J., Bordage, G., Gierl, M. J. & Yudkowsky, R. 2014. Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an Internal Medicine residency OSCE. *Advances in Health Sciences Education*, 19, 497-506.

- Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E. & Perrier, L. 2006. Accuracy of physician self-assessment compared with observed measures of competence - A systematic review. *Jama-Journal of the American Medical Association*, 296, 1094-1102.
- De Almeida, C. E., Rodriguez, M., Vianello, E., Ferreira, I. H. & Sibata, C. 2002. An anthropomorphic phantom for quality assurance and training in gynaecological brachytherapy. *Radiotherapy and Oncology*, 63, 75-81.
- De Bruijn-Smolters, M., Timmers, C. F., Gawke, J. C. L., Schoonman, W. & Born, M. P. 2014. Effective self-regulatory processes in higher education: research findings and future directions. A systematic review. *Studies in Higher Education*, 41, 139-158.
- De Jong, T. 2009. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, 38, 105-134.
- Devine, L. A., Mcgaghie, W. C. & Issenberg, B. 2019. Simulations in assessment. In: Yudkowsky, R., Park, Y. S. & Downing, S. M. (eds.) *Assessment in Health Professions Education*. 2nd ed. New York, NY: Routledge.
- Dewey, J. 1925. *Experience and nature*, Whitefish, MT, Kessinger.
- Dimopoulos, J. C. A., Vos, V. D., Berger, D., Petric, P., Dumas, I., Kirisits, C., . . . Pötter, R. 2009. Inter-observer comparison of target delineation for MRI-assisted cervical cancer brachytherapy: Application of the GYN GEC-ESTRO recommendations. *Radiotherapy and Oncology*, 91, 166-172.
- Disessa, A. A. & Cobb, P. 2004. Ontological innovation and the role of theory in design experiments. *The journal of the learning sciences*, 13, 77-103.
- Dolmans, D. H. & Tigelaar, D. 2012. Building bridges between theory and practice in medical education using a design-based research approach: AMEE Guide No. 60. *Medical Teacher*, 34, 1-10.
- Donaldson, M. S., Corrigan, J. M. & Kohn, L. T. 2000. To err is human: building a safer health system. National Academies Press.
- Downing, S. M. 2003. Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37, 830-837.
- Downing, S. M. 2004. Reliability: on the reproducibility of assessment data. *Medical Education*, 38, 1006-1012.
- Downing, S. M. & Haladyna, T. M. 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38, 327-333.
- Ducreux, M., Cuhna, A. S., Caramella, C., Hollebecque, A., Burtin, P., Goere, D., . . . Committee, E. G. 2015. Cancer of the pancreas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 26 Suppl 5, v56-68.
- Ebbinghaus, H. 1885. *Über das Gedächtnis*, Leipzig, Dunker.
- Eisbruch, A., Marsh, L. H., Dawson, L. A., Bradford, C. R., Teknos, T. N., Chepeha, D. B., . . . Wolf, G. T. 2004. Recurrences near base of skull after IMRT for head-and-neck cancer: implications for target delineation in high neck and for parotid gland sparing. *International Journal of Radiation Oncology, Biology, Physics*, 59, 28-42.

- Eminowicz, G., Hall-Craggs, M., Diez, P. & McCormack, M. 2016a. Improving target volume delineation in intact cervical carcinoma: Literature review and step-by-step pictorial atlas to aid contouring. *Practical Radiation Oncology*, 6, e203-e213.
- Eminowicz, G. & McCormack, M. 2015. Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. *Radiotherapy and Oncology*, 117, 542-547.
- Eminowicz, G., Rompokos, V., Stacey, C. & McCormack, M. 2016b. The dosimetric impact of target volume delineation variation for cervical cancer radiotherapy. *Radiotherapy and Oncology*, 120, 493-499.
- Eppich, W. & Cheng, A. 2015. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. *Simul Healthc*, 10, 106-15.
- Ericsson, K. A. 2007a. Deliberate practice and the modifiability of body and mind: toward a science of the structure and acquisition of expert and elite performance. *International Journal of Sport Psychology*, 38, 4-34.
- Ericsson, K. A. 2007b. An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education*, 41, 1124-30.
- Ericsson, K. A. 2015. Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach with deliberate practice. *Academic Medicine*, 90, 1471-1486.
- Ericsson, K. A. & Charness, N. 1994. Expert performance: Its structure and acquisition. *American Psychologist*, 49, 725.
- Ericsson, K. A., Krampe, R. T. & Teschmer, C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.
- Ericsson, K. A. & Lehmann, A. C. 1996. Expert and exceptional performance: evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, 47, 273-305.
- Ericsson, K. A. & Simon, H. A. 1980. Verbal reports as data. *Psychological Review*, 87, 215.
- Eriksen, J. G., Beavis, A. W., Coffey, M. A., Leer, J. W., Magrini, S. M., Benstead, K., . . . Potter, R. 2012. The updated ESTRO core curricula 2011 for clinicians, medical physicists and RTTs in radiotherapy/radiation oncology. *Radiotherapy and Oncology*, 103, 103-8.
- Eva, K. W. 2008. On the limits of systematicity. *Medical Education*, 42, 852-3.
- Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E. & Regehr, G. 2016. Towards a program of assessment for health professionals: from training into practice. *Advances in Health Sciences Education*, 21, 897-913.
- Evans, E., Jones, G., Rackley, T., Maggs, R., Radhakrishna, G., Mukherjee, S., . . . Gwynne, S. 2018. PO-0769: NeoSCOPE RTTQA: pre-accrual and on-trial review of all patients in a UK oesophageal RT trial. *Radiotherapy and Oncology*, 127.
- Evans, E., Piazzese, C., Spezi, E., Staffurth, J. & Gwynne, S. 2019a. ARENA: Improving training in target volume delineation for radiotherapy. *Radiotherapy and Oncology*, 133, S896-S897.

- Evans, E., Radhakrishna, G., Gilson, D., Hoskin, P., Miles, E., Yuille, F., . . . Gwynne, S. 2019b. Target Volume Delineation Training for Clinical Oncology Trainees: The Role of ARENA and COPP. *Clinical Oncology (Royal College of Radiologists)*, 31, 341-343.
- Feilzer, M. Y. 2009. Doing Mixed Methods Research Pragmatically: Implications for the Rediscovery of Pragmatism as a Research Paradigm. *Journal of Mixed Methods Research*, 4, 6-16.
- Fokas, E., Clifford, C., Spezi, E., Joseph, G., Branagan, J., Hurt, C., . . . Mukherjee, S. 2015. Comparison of investigator-delineated gross tumor volumes and quality assurance in pancreatic cancer: Analysis of the pretrial benchmark case for the SCALOP trial. *Radiotherapy and Oncology*, 117, 432-7.
- Fraefel, U. Professionalization of pre-service teachers through university-school partnerships. Conference Proceedings of WERA Focal Meeting, Edinburgh, 2014.
- Gaba, D. M., Howard, S. K., Fish, K. J., Smith, B. E. & Sowb, Y. A. 2016. Simulation-Based Training in Anesthesia Crisis Resource Management (ACRM): A Decade of Experience. *Simulation & Gaming*, 32, 175-193.
- Gardner, H. 1995. Why would anyone become an expert? *American Psychologist*, 50, 802-803.
- Gautam, A., Weiss, E., Williamson, J., Ford, J., Sleeman, W., Jan, N., . . . Murphy, M. 2013. SU-C-WAB-03: Assessing the Correlation Between Quantitative Measures of Contour Variability and Physician's Qualitative Measure for Clinical Usefulness of Auto-Segmentation in Prostate Cancer Radiotherapy. *Medical Physics*, 40, 90-90.
- Gentner, D. & Stevens, A. L. 2014. *Mental models*, East Sussex, England, Psychology Press.
- George, T. 2020. *Hermeneutics* [Online]. Stanford University,. [Accessed 11/01/2020].
- Gillespie, E. F., Panjwani, N., Golden, D. W., Gunther, J., Chapman, T. R., Brower, J. V., . . . Murphy, J. D. 2017. Multi-institutional Randomized Trial Testing the Utility of an Interactive Three-dimensional Contouring Atlas Among Radiation Oncology Residents. *International Journal of Radiation Oncology, Biology, Physics*, 98, 547-554.
- Gostlow, H., Marlow, N., Babidge, W. & Maddern, G. 2017. Systematic Review of Voluntary Participation in Simulation-Based Laparoscopic Skills Training: Motivators and Barriers for Surgical Trainee Attendance. *Journal of Surgical Education*, 74, 306-318.
- Greenhalgh, T., A'court, C. & Shaw, S. 2017. Understanding heart failure; explaining telehealth - a hermeneutic systematic review. *BMC Cardiovascular Disorders*, 17, 156.
- Greenhalgh, T. & Peacock, R. 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ*, 331, 1064-5.
- Greenhalgh, T., Thorne, S. & Malterud, K. 2018. Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, 48, e12931.
- Grégoire, V., Ang, K., Budach, W., Grau, C., Hamoir, M., Langendijk, J. A., . . . Lengele, B. 2014a. *Delineation of the neck node levels for head and neck tumors - online atlas* [Online]. Available: <https://www.nrgoncology.org/Portals/0/Scientific%20Program/CIRO/Atlases/HNC%20new%20atlas.pdf?ver=2020-08-19-112410-653> [Accessed].

- Grégoire, V., Ang, K., Budach, W., Grau, C., Hamoir, M., Langendijk, J. A., . . . Lengele, B. 2014b. Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiotherapy and Oncology*, 110, 172-181.
- Group, J. C. O., Toita, T., Ohno, T., Kaneyasu, Y., Uno, T., Yoshimura, R., . . . Hiraoka, M. 2010. A consensus-based guideline defining the clinical target volume for pelvic lymph nodes in external beam radiotherapy for uterine cervical cancer. *Japanese Journal of Clinical Oncology*, 40, 456-63.
- Grover, S. C., Scaffidi, M. A., Khan, R., Garg, A., Al-Mazroui, A., Alomani, T., . . . Walsh, C. M. 2017. Progressive learning in endoscopy simulation training improves clinical performance: a blinded randomized trial. *Gastrointestinal Endoscopy*, 86, 881-889.
- Guadagnoli, M., Morin, M. P. & Dubrowski, A. 2012. The application of the challenge point framework in medical education. *Medical Education*, 46, 447-53.
- Guadagnoli, M. A., Dornier, L. A. & Tandy, R. D. 1996. Optimal Length for Summary Knowledge of Results: The Influence of Task-Related Experience and Complexity. *Research Quarterly for Exercise and Sport*, 67, 239-248.
- Gwynne, S., Gilson, D., Dickson, J., Mcaleer, S. & Radhakrishna, G. 2017. Evaluating Target Volume Delineation in the Era of Precision Radiotherapy: FRCR, Revalidation and Beyond. *Clinical Oncology*, 29, 436-438.
- Gwynne, S., Spezi, E., Sebag-Montefiore, D., Mukherjee, S., Miles, E., Conibear, J., . . . Imaging, S. 2013. Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case. *British Journal of Radiology*, 86, 20120398.
- Haie-Meder, C., Pötter, R., Van Limbergen, E., Briot, E., De Brabandere, M., Dimopoulos, J., . . . Wachter-Gerstner, N. 2005. Recommendations from Gynaecological (GYN) GEC-ESTRO Working Group (I): Concepts and terms in 3D image based 3D treatment planning in cervix cancer brachytherapy with emphasis on MRI assessment of GTV and CTV. *Radiotherapy and Oncology*, 74, 235-245.
- Haig, A., Dozier, M., Liu, D., Mckendree, J., Roper, T. & Selai, C. 2005. METRO taxonomy - progress report on assessment. *Medical Teacher*, 27, 155-7.
- Haig, A., Ellaway, R., Dozier, M., Liu, D. & Mckendree, J. 2004. METRO--the creation of a taxonomy for medical education. *Health Info Libr J*, 21, 211-9.
- Haji, F. A., Cheung, J. J., Woods, N., Regehr, G., De Ribaupierre, S. & Dubrowski, A. 2016. Thrive or overload? The effect of task complexity on novices' simulation-based learning. *Medical Education*, 50, 955-68.
- Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B. & Cook, D. A. 2014. Reconsidering fidelity in simulation-based training. *Academic Medicine*, 89, 387-92.
- Hanna, G. G., Hounsell, A. R. & O'sullivan, J. M. 2010. Geometrical Analysis of Radiotherapy Target Volume Delineation: A Systematic Review of Reported Comparison Methods. *Clinical Oncology*, 22, 515-525.
- Hata, M., Miyagi, E., Koike, I., Numazaki, R., Asai-Sato, M., Kasuya, T., . . . Inoue, T. 2015. Radiation therapy for para-aortic lymph node metastasis from uterine cervical cancer. *International Journal of Gynecological Cancer*, 25, 189-189.

- Hatala, R., Cook, D. A., Zendejas, B., Hamstra, S. J. & Brydges, R. 2014. Feedback for simulation-based procedural skills training: a meta-analysis and critical narrative synthesis. *Advances in Health Sciences Educational Theory and Practice*, 19, 251-72.
- Hatala, R., Kassen, B. O., Nishikawa, J., Cole, G. & Issenberg, S. B. 2005. Incorporating simulation technology in a canadian internal medicine specialty examination: a descriptive report. *Academic Medicine*, 80, 554-6.
- Hattie, J. 1999. Influences on student learning. *Inaugural professorial address, University of Auckland, New Zealand*.
- Hattie, J. 2009. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*, Abingdon, Oxfordshire, Routledge.
- Hattie, J. & Timperley, H. 2007. The Power of Feedback. *Review of Educational Research*, 77, 81-112.
- Hege, I., Kononowicz, A. A. & Adler, M. 2017. A Clinical Reasoning Tool for Virtual Patients: Design-Based Research Study. *JMIR Med Educ*, 3, e21.
- Heidegger, M. 1929. *Sein und Zeit*, Niemeyer.
- Hellebust, T. P., Tanderup, K., Lervag, C., Fidarova, E., Berger, D., Malinen, E., . . . Petric, P. 2013. Dosimetric impact of interobserver variability in MRI-based delineation for cervical cancer brachytherapy. *Radiotherapy and Oncology*, 107, 13-9.
- Henriksen, K., Rodrick, D., Grace, E. N. & Brady, P. J. 2018. Challenges in Health Care Simulation: Are We Learning Anything New? *Academic Medicine*, 93, 705-708.
- Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J. & Welch, V. A. 2019. *Cochrane handbook for systematic reviews of interventions*, John Wiley & Sons.
- Higgins, M., Madan, C. & Patel, R. 2020. Development and decay of procedural skills in surgery: A systematic review of the effectiveness of simulation-based medical education interventions. *Surgeon*.
- Huq, M. S., Fraass, B. A., Dunscombe, P. B., Gibbons, J. P., Jr., Ibbott, G. S., Mundt, A. J., . . . Yorke, E. D. 2016. The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. *Medical Physics*, 43, 4209.
- Icru. ICRU Report No. 50. Prescribing, recording and reporting photon beam therapy. 1993. International Commission on Radiation Units and Measurements, Washington, DC.
- Icru 1999. ICRU Report No. 62: Prescribing, Recording, and Reporting Photon Beam Therapy (Supplement to ICRU Report 50).
- Icru 2013. ICRU Report No. 89: Prescribing, Recording, and Reporting Brachytherapy for Cancer of the Cervix. *Journal of the ICRU*, 13.
- Ims Global. 2020. *Learning tools Interoperability 1.3* [Online]. Available: <http://www.imsglobal.org/activity/learning-tools-interoperability> [Accessed 01/02/2021].
- Iso 2010. Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. *International Standards Organisation*, 2010, 1-32.

- Issenberg, S. B., Mcgaghie, W. C., Hart, I. R., Mayer, J. W., Felner, J. M., Petrusa, E. R., . . . Ewy, G. A. 1999. Simulation technology for health care professional skills training and assessment. *JAMA*, 282, 861-6.
- Issenberg, S. B., Mcgaghie, W. C., Petrusa, E. R., Lee Gordon, D. & Scalese, R. J. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher*, 27, 10-28.
- Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., . . . Oxman, A. D. 2012. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews*.
- Jadon, R., Pembroke, C. A., Hanna, C. L., Palaniappan, N., Evans, M., Cleves, A. E. & Staffurth, J. 2014. A systematic review of organ motion and image-guided strategies in external beam radiotherapy for cervical cancer. *Clinical Oncology (Royal College of Radiologists)*, 26, 185-96.
- Jaffray, D. A. & Gospodarowicz, M. K. 2015. Radiation Therapy for Cancer. *Disease Control Priorities, Third Edition (Volume 3): Cancer*. The World Bank.
- Jameson, M. G., Holloway, L. C., Vial, P. J., Vinod, S. K. & Metcalfe, P. E. 2010. A review of methods of analysis in contouring studies for radiation oncology. *Journal of Medical Imaging and Radiation Oncology*, 54, 401-10.
- Jameson, M. G., Kumar, S., Vinod, S. K., Metcalfe, P. E. & Holloway, L. C. 2014. Correlation of contouring variation with modeled outcome for conformal non-small cell lung cancer radiotherapy. *Radiotherapy and Oncology*, 112, 332-336.
- Jensen, N. B. K., Potter, R., Spampinato, S., Fokdal, L. U., Chargari, C., Lindegaard, J. C., . . . Group, E. C. 2021. Dose-Volume Effects and Risk Factors for Late Diarrhea in Cervix Cancer Patients After Radiochemotherapy With Image Guided Adaptive Brachytherapy in the EMBRACE I Study. *International Journal of Radiation Oncology, Biology, Physics*, 109, 688-700.
- Joo, J. H., Kim, Y. S., Cho, B. C., Jeong, C. Y., Park, W., Kim, H. J., . . . Kim, J. Y. 2017. Variability in target delineation of cervical carcinoma: A Korean radiation oncology group study (KROG 15-06). *PloS One*, 12, e0173476.
- Jordan, K. 2014. Initial trends in enrolment and completion of massive open online courses. *Int Rev Res Open Dist Learn*, 15.
- Kalyuga, S. 2011. Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, 23, 1-19.
- Kalyuga, S., Ayres, P., Chandler, P. & Sweller, J. 2003. The expertise reversal effect. *Educational Psychologist*, 38, 23-31.
- Kane, M. T. 1992. The Assessment of Professional Competence. *Evaluation and the Health Professions*, 15, 163-182.
- Keenan, L. G., Rock, K., Azmi, A., Salib, O., Gillham, C. & Mcardle, O. 2018. An atlas to aid delineation of para-aortic lymph node region in cervical cancer: Design and validation of contouring guidelines. *Radiotherapy and Oncology*, 127, 417-422.
- Kelly, A. 2004. Design Research in Education: Yes, but is it Methodological? *Journal of the Learning Sciences*, 13, 115-128.

- Kelly, C., Thirion, P., Grimley, S., Dimopoulos, J. & Potter, R. 2006. Quantification of interobserver variation in delineation of target volumes using the GEC-ESTRO recommendations for MRI based brachytherapy of the cervix. *Radiotherapy and Oncology*, 81.
- Kettley, N. 2010. *Theory building in educational research*, Bloomsbury Publishing.
- Kim, R. Y., McGinnis, L. S., Spencer, S. A., Meredith, R. F., Jennelle, R. L. S. & Salter, M. M. 1995. Conventional four-field pelvic radiotherapy technique without computed tomography-treatment planning in cancer of the cervix: Potential geographic miss and its impact on pelvic control. *International Journal of Radiation Oncology, Biology, Physics*, 31, 109-112.
- Kirisits, C., Federico, M., Nkiwane, K., Fidarova, E., Jurgenliemk-Schulz, I., De Leeuw, A., . . . Tanderup, K. 2015. Quality assurance in MR image guided adaptive brachytherapy for cervical cancer: Final results of the EMBRACE study dummy run. *Radiotherapy and Oncology*, 117, 548-554.
- Kirkpatrick, D. L. 1967. Evaluation of training. *Training and Development Handbook*. New York: McGraw-Hill.
- Kirschner, P. A., Sweller, J. & Clark, R. E. 2006. Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching *Educational Psychologist*, 41, 87-98.
- Kirschner, P. A. & Van Merriënboer, J. J. G. 2013. Do Learners Really Know Best? Urban Legends in Education. *Educational Psychologist*, 48, 169-183.
- Kluger, A. N. & Denisi, A. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254.
- Ko, J. & Sammons, P. 2013. *Effective Teaching: A Review of Research and Evidence*, ERIC.
- Konopasek, L., Norcini, J. & Krupat, E. 2016. Focusing on the Formative: Building an Assessment System Aimed at Student Growth and Development. *Academic Medicine*, 91, 1492-1497.
- Kruger, J. & Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121-1134.
- Lammers, R. L. 2008. Learning and Retention Rates after Training in Posterior Epistaxis Management. *Academic Emergency Medicine*, 15, 1181-1189.
- Larsen, D. P., Butler, A. C. & Roediger, H. L. 2008. Test-enhanced learning in medical education. *Medical Education*, 42, 959-966.
- Lefroy, J., Watling, C., Teunissen, P. W. & Brand, P. 2015. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. *Perspect Med Educ*, 4, 284-99.
- Lewis, C. 1982. Using the "think aloud" method in cognitive interface design. *New York: IBM*.
- Lewis, J. R. 2014. Usability: Lessons Learned. and Yet to Be Learned. *International Journal of Human-Computer Interaction*, 30, 663-684.
- Lewis, J. R. 2018. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction*, 34, 577-590.

- Lim, K., Small, W., Jr., Portelance, L., Creutzberg, C., Jurgenliemk-Schulz, I. M., Mundt, A., . . . Gyn, I. C. 2011. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. *International Journal of Radiation Oncology, Biology, Physics*, 79, 348-55.
- Lim, T. Y., Gillespie, E., Murphy, J. & Moore, K. L. 2019. Clinically Oriented Contour Evaluation Using Dosimetric Indices Generated From Automated Knowledge-Based Planning. *International Journal of Radiation Oncology, Biology, Physics*, 103, 1251-1260.
- Lin, D., Lapen, K., Sherer, M. V., Kantor, J., Zhang, Z., Boyce, L. M., . . . Gillespie, E. F. 2020. A Systematic Review of Contouring Guidelines in Radiation Oncology: Analysis of Frequency, Methodology, and Delivery of Consensus Recommendations. *International Journal of Radiation Oncology, Biology, Physics*, 107, 827-835.
- Lineberry, M., Soo Park, Y., Cook, D. A. & Yudkowsky, R. 2015. Making the case for mastery learning assessments: key issues in validation and justification. *Academic Medicine*, 90, 1445-50.
- Lioce, L., Loperalto, J., Downing, D., Chang, T. P., Robertson, J. M., Anderson, M. & Diaz, D. A. 2020. *Healthcare Simulation Dictionary - Second Edition*, Rockville, MD, Agency for Healthcare Research and Quality.
- Lobefalo, F., Bignardi, M., Reggiori, G., Tozzi, A., Tomatis, S., Alongi, F., . . . Mancosu, P. 2013. Dosimetric impact of inter-observer variability for 3D conformal radiotherapy and volumetric modulated arc therapy: the rectal tumor target definition case. *Radiation Oncology (London, England)*, 8, 176.
- Looyestyn, J., Kernot, J., Boshoff, K., Ryan, J., Edney, S. & Maher, C. 2017. Does gamification increase engagement with online programs? A systematic review. *PloS One*, 12, e0173403.
- Macnamara, B. N., Hambrick, D. Z. & Oswald, F. L. 2014. Deliberate practice and performance in music, games, sports, education, and professions: a meta-analysis. *Psychological Science*, 25, 1608-18.
- Macnamara, B. N. & Maitra, M. 2019. The role of deliberate practice in expert performance: revisiting Ericsson, Krampe & Tesch-Romer (1993). *R Soc Open Sci*, 6, 190327.
- Manea, E., Escande, A., Bockel, S., Khettab, M., Dumas, I., Lazarescu, I., . . . Chargari, C. 2018. Risk of Late Urinary Complications Following Image Guided Adaptive Brachytherapy for Locally Advanced Cervical Cancer: Refining Bladder Dose-Volume Parameters. *International Journal of Radiation Oncology, Biology, Physics*, 101, 411-420.
- Marks, L. B., Yorke, E. D., Jackson, A., Ten Haken, R. K., Constine, L. S., Eisbruch, A., . . . Deasy, J. O. 2010. Use of Normal Tissue Complication Probability Models in the Clinic. *International Journal of Radiation Oncology Biology Physics*, 76.
- Marth, C., Landoni, F., Mahner, S., McCormack, M., Gonzalez-Martin, A., Colombo, N. & Committee, E. G. 2017. Cervical cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 28, iv72-iv83.
- Maxwell, J. A. 2006. Literature reviews of, and for, educational research: A commentary on Boote and Beile's "Scholars before Researchers". *Educational Researcher*, 35, 28-31.
- Mccarroll, R. E., Beadle, B. M., Balter, P. A., Burger, H., Cardenas, C. E., Dalvie, S., . . . Yang, J. 2018. Retrospective Validation and Clinical Implementation of Automated Contouring of Organs at

- Risk in the Head and Neck: A Step Toward Automated Radiation Treatment Planning for Low- and Middle-Income Countries. *Journal of Global Oncology*, 4, 1-11.
- McGaghie, W. C. 2015. Mastery learning: it is time for medical education to join the 21st century. *Academic Medicine*, 90, 1438-41.
- McGaghie, W. C., Issenberg, S. B., Barsuk, J. H. & Wayne, D. B. 2014. A critical review of simulation-based mastery learning with translational outcomes. *Medical Education*, 48, 375-85.
- McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H. & Wayne, D. B. 2011. Does Simulation-Based Medical Education With Deliberate Practice Yield Better Results Than Traditional Clinical Education? A Meta-Analytic Comparative Review of the Evidence. *Academic Medicine*, 86, 706-711.
- McGaghie, W. C., Issenberg, S. B., Petrusa, E. R. & Scalese, R. J. 2010. A critical review of simulation-based medical education research: 2003-2009. *Medical Education*, 44, 50-63.
- McGaghie, W. C., Siddall, V. J., Mazmanian, P. E., Myers, J., American College of Chest Physicians, H. & Science Policy, C. 2009. Lessons for continuing medical education from simulation research in undergraduate and graduate medical education: effectiveness of continuing medical education: American College of Chest Physicians Evidence-Based Educational Guidelines. *Chest*, 135, 62S-68S.
- McKenney, S. & Reeves, T. C. 2020. Educational design research: Portraying, conducting, and enhancing productive scholarship. *Medical Education*.
- McKinley, D. W. & Norcini, J. J. 2013. How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*, 36, 97-110.
- McLaughlin, P. W., Evans, C., Feng, M. & Narayana, V. 2010. Radiographic and Anatomic Basis for Prostate Contouring Errors and Methods to Improve Prostate Contouring Accuracy. *International Journal of Radiation Oncology Biology Physics*, 76, 369-378.
- McMillan, W. 2015. Theory in healthcare education research: the importance of worldview. In: Cleland, J. & Durning, S. (eds.) *Researching Medical Education*. Wiley Blackwell.
- Mehay, R. & Burns, R. 2009. Miller's pyramid/prism of clinical competence. In: Mehay, R. (ed.) *The essential handbook for GP training and education*. London, UK: Radcliffe Publishing.
- Melidis, C., Bosch, W. R., Izewska, J., Fidarova, E., Zubizarreta, E., Ishikura, S., . . . Hurkmans, C. W. 2014. Radiation therapy quality assurance in clinical trials--Global Harmonisation Group. *Radiotherapy and Oncology*, 111, 327-9.
- Mercieca, S., Pan, S., Belderbos, J., Salem, A., Tenant, S., Aznar, M. C., . . . Van Herk, M. 2020. Impact of Peer Review in Reducing Uncertainty in the Definition of the Lung Target Volume Among Trainee Oncologists. *Clinical Oncology (Royal College of Radiologists)*, 32, 363-372.
- Meyer, H. S., Durning, S. J., Sklar, D. P. & Maggio, L. A. 2018. Making the First Cut: An Analysis of Academic Medicine Editors' Reasons for Not Sending Manuscripts Out for External Peer Review. *Academic Medicine*, 93, 464-470.
- Michalski, J. M., Lawton, C., El Naqa, I., Ritter, M., O'meara, E., Seider, M. J., . . . Ménard, C. 2010. Development of RTOG Consensus Guidelines for the Definition of the Clinical Target Volume for Postoperative Conformal Radiation Therapy for Prostate Cancer. *International Journal of Radiation Oncology Biology Physics*, 76, 361-368.

- Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Miller, G. E. 1990. The assessment of clinical skills/competence/performance. *Academic Medicine*, 65, S63-7.
- Montero, P. N., Acker, C. E., Heniford, B. T. & Stefanidis, D. 2011. Single incision laparoscopic surgery (SILS) is associated with poorer performance and increased surgeon workload compared with standard laparoscopy. *American Surgeon*, 77, 73-7.
- Moore, M. J. & Bennett, C. L. 1995. The learning curve for laparoscopic cholecystectomy. *The American Journal of Surgery*, 170, 55-59.
- Morarji, K., Fowler, A., Vinod, S. K., Ho Shon, I. & Laurence, J. M. 2012. Impact of FDG-PET on lung cancer delineation for radiotherapy. *Journal of Medical Imaging and Radiation Oncology*, 56, 195-203.
- Moreno, R. 2009. Cognitive load theory: more food for thought. *Instructional Science*, 38, 135-141.
- Morrison, K. 2018. Coding and content analysis. *Research methods in education*. Routledge.
- Muijs, C. T., Schreurs, L. M., Busz, D. M., Beukema, J. C., Van Der Borden, A. J., Pruim, J., . . . Langendijk, J. A. 2009. Consequences of additional use of PET information for target volume delineation and radiotherapy dose distribution for esophageal cancer. *Radiotherapy and Oncology*, 93, 447-53.
- Murphy, G., Groeger, J. A. & Greene, C. M. 2016. Twenty years of load theory-Where are we now, and where should we go next? *Psychonomic Bulletin & Review*, 23, 1316-1340.
- Murphy, J. G. & Gillespie, E. F. 2019. *NIH (US) Grant Award - Interactive contouring simulation (iContour) to improve outcomes of cancer patients treated with radiation therapy* [Online]. Available: <https://grantome.com/grant/NIH/R18-HS026881-01> [Accessed 13/03/2021].
- Naismith, L. M. & Cavalcanti, R. B. 2015. Validity of cognitive load measures in simulation-based training: A systematic review. *Academic Medicine*, 90, S24-S35.
- Nasca, T. J., Philibert, I., Brigham, T. & Flynn, T. C. 2012. The next GME accreditation system--rationale and benefits. *New England Journal of Medicine*, 366, 1051-6.
- National Institute for Health Research (Uk). 2020. *Clinical Education Incubator* [Online]. Available: <https://www.nihr.ac.uk/documents/clinical-education-incubator/24887> [Accessed 24/03/2021].
- Neoptolemos, J. P., Stocken, D. D., Friess, H., Bassi, C., Dunn, J. A., Hickey, H., . . . European Study Group for Pancreatic, C. 2004. A randomized trial of chemoradiotherapy and chemotherapy after resection of pancreatic cancer. *New England Journal of Medicine*, 350, 1200-10.
- Nielsen, J. & Landauer, T. K. 1993. A mathematical model of the finding of usability problems. *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 206-213.
- Nieveen, N. & Folmer, E. 2013. Formative evaluation in educational design research. *Design Research*, 153, 152-169.

- Nisbett, R. E. & Wilson, T. D. 1977. The halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35, 250.
- Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., . . . Roberts, T. 2011. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33, 206-14.
- Norman, G., Dore, K. & Grierson, L. 2012. The minimal relationship between simulation fidelity and transfer of learning. *Medical Education*, 46, 636-647.
- Norman, G. R., Grierson, L. E. M., Sherbino, J., Hamstra, S. J., Schmidt, H. G. & Mamede, S. 2018. Expertise in Medicine and Surgery. In: Williams, A. M., Kozbelt, A., Ericsson, K. A. & Hoffman, R. R.(eds.) *The Cambridge Handbook of Expertise and Expert Performance*. 2 ed. Cambridge: Cambridge University Press.
- Nutting, C. M., Morden, J. P., Harrington, K. J., Urbano, T. G., Bhide, S. A., Clark, C., . . . Hall, E. 2011. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): A phase 3 multicentre randomised controlled trial. *The Lancet Oncology*, 12, 127-136.
- Offersen, B. V., Boersma, L. J., Kirkove, C., Hol, S., Aznar, M. C., Biete Sola, A., . . . Poortmans, P. 2015. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and Oncology*, 114, 3-10.
- Ohri, N., Shen, X., Dicker, A. P., Doyle, L. A., Harrison, A. S. & Showalter, T. N. 2013. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. *Journal of the National Cancer Institute*, 105, 387-93.
- Olszewski, A. E. & Wolbrink, T. A. 2017. Serious Gaming in Medical Education: A Proposed Structured Framework for Game Development. *Simul Healthc*, 12, 240-253.
- Onwuegbuzie, A. J. & Leech, N. L. 2005. On Becoming a Pragmatic Researcher: The Importance of Combining Quantitative and Qualitative Research Methodologies. *International Journal of Social Research Methodology*, 8, 375-387.
- Open Science Collaboration 2015. Estimating the reproducibility of psychological science. *Science*, 349, aac4716.
- Ouyang, Y., Wang, Y., Chen, K., Cao, X. & Zeng, Y. 2017. Clinical outcome of extended-field irradiation vs. pelvic irradiation using intensity-modulated radiotherapy for cervical cancer. *Oncology Letters*, 14, 7069-7076.
- Panadero, E., Jonsson, A. & Botella, J. 2017. Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98.
- Pano, B., Sebastia, C., Ripoll, E., Paredes, P., Salvador, R., Bunesch, L. & Nicolau, C. 2015. Pathways of lymphatic spread in gynecologic malignancies. *Radiographics*, 35, 916-45.
- Papinczak, T., Young, L., Groves, M. & Haynes, M. 2008. Effects of a metacognitive intervention on students' approaches to learning and self-efficacy in a first year medical course. *Adv Health Sci Educ Theory Pract*, 13, 213-32.
- Park, B. & Brünken, R. 2015. The Rhythm Method: A New Method for Measuring Cognitive Load - An Experimental Dual-Task Study. *Applied Cognitive Psychology*, 29, 232-243.

- Pell, G., Fuller, R., Homer, M. & Roberts, T. 2013. Advancing the objective structured clinical examination: sequential testing in theory and practice. *Medical Education*, 47, 569-77.
- Peters, L. J., O'sullivan, B., Giralt, J., Fitzgerald, T. J., Trotti, A., Bernier, J., . . . Rischin, D. 2010. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *Journal of Clinical Oncology*, 28, 2996-3001.
- Petric, P., Dimopoulos, J., Kirisits, C., Berger, D., Hudej, R. & Potter, R. 2008. Inter- and intraobserver variation in HR-CTV contouring: intercomparison of transverse and paratransverse image orientation in 3D-MRI assisted cervix cancer brachytherapy. *Radiotherapy and Oncology*, 89, 164-71.
- Petric, P., Hudej, R., Rogelj, P., Blas, M., Tanderup, K., Fidarova, E., . . . Hellebust, T. P. 2013. Uncertainties of target volume delineation in MRI guided adaptive brachytherapy of cervix cancer: a multi-institutional study. *Radiotherapy and Oncology*, 107, 6-12.
- Pilcher, J., Goodall, H., Jensen, C., Huwe, V., Jewell, C., Reynolds, R. & Karlsen, K. A. 2012. Special focus on simulation: educational strategies in the NICU: simulation-based learning: it's not just for NRP. *Neonatal Network*, 31, 281-7.
- Plomp, T. & Nieveen, N. 2013. *Educational design research: an introduction*, Enschede, the Netherlands, SLO.
- Pötter, R., Lindegaard, J., Kirisits, C., Juergenliemk-Schulz, I., Leeuw, A. D., Fortin, I., . . . Tan, L. T. 2016a. *EMBRACE-II Study Protocol* [Online]. Available: <https://www.embracestudy.dk/UserUpload/PublicDocuments/Docs/EMBRACE II title page and total protocol v15 151015.pdf> [Accessed 01/12/2020].
- Pötter, R., Lindegaard, J., Kirisits, C., Juergenliemk-Schulz, I., Leeuw, A. D., Fortin, I., . . . Tan, L. T. 2016b. *EMBRACE-II Study Website* [Online]. Available: <https://embracestudy.dk> [Accessed 01/12/2020].
- Pötter, R., Tanderup, K., Kirisits, C., De Leeuw, A., Kirchheiner, K., Nout, R., . . . Jürgenliemk-Schulz, I. 2018. The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. *Clin Transl Radiat Oncol*, 9, 48-60.
- Pötter, R., Tanderup, K., Schmid, M., Haie-Meder, C., Fokdal, L. U., Sturdza, A., . . . Embrace, C. G. 2020. OC-0437: MRI guided adaptive brachytherapy in locally advanced cervical cancer: overall results of EMBRACE I. *Radiotherapy and Oncology*, 152, S240-S241.
- Radford, P. D., Derbyshire, L. F., Shalhoub, J., Fitzgerald, J. E. F. & Council of the Association of Surgeons in Training 2015. Publication of surgeon specific outcome data: a review of implementation, controversies and the potential impact on surgical training. *International Journal of Surgery (London, England)*, 13, 211-216.
- Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E. & Wright, S. M. 2007. Association between funding and quality of published medical education research. *Jama-Journal of the American Medical Association*, 298, 1002-1009.
- Richardson, M., Abraham, C. & Bond, R. 2012. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin*, 138, 353.

- Riegel, A. C., Berson, A. M., Destian, S., Ng, T., Tena, L. B., Mitnick, R. J. & Wong, P. S. 2006. Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. *International Journal of Radiation Oncology, Biology, Physics*, 65, 726-32.
- Rivin Del Campo, E., Rivera, S., Martínez-Paredes, M., Hupé, P., Slocker Escarpa, A., Borget, I., . . . Deutsch, E. 2017. Assessment of the novel online delineation workshop dummy run approach using FALCON within a European multicentre trial in cervical cancer (RAIDs). *Radiotherapy and Oncology*, 124, 130-138.
- Roeske, J. C., Forman, J. D., Mesina, C. F., He, T., Pelizzari, C. A., Fontenla, E., . . . Chen, G. T. 1995. Evaluation of changes in the size and location of the prostate, seminal vesicles, bladder, and rectum during a course of external beam radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, 33, 1321-9.
- Rohrer, D. & Pashler, H. 2010. Recent Research on Human Learning Challenges Conventional Instructional Strategies. *Educational Researcher*, 39, 406-412.
- Romanchuk, K. 2004. The Effect of Limiting Residents' Work Hours on Their Surgical Training: A Canadian Perspective. *Academic Medicine*, 79, 384-385.
- Rooney, M. K., Zhu, F., Gillespie, E. F., Gunther, J. R., Mckillip, R. P., Lineberry, M., . . . Golden, D. W. 2018. Simulation as More Than a Treatment-Planning Tool: A Systematic Review of the Literature on Radiation Oncology Simulation-Based Medical Education. *International Journal of Radiation Oncology, Biology, Physics*, 102, 257-283.
- Royal College of Radiologists. 2020a. *Clinical Oncology Specialty Training Curriculum* [Online]. Available: https://www.rcr.ac.uk/sites/default/files/clinical_oncology_curriculum_2021.pdf [Accessed].
- Royal College of Radiologists. 2020b. *Imaging for oncology trainees* [Online]. [Accessed 31/12/2020].
- Royal College of Surgeons of England. 2014. *Surgical Outcomes: driving up standards of care for patients through publishing surgeons outcomes data* [Online]. Available: <https://www.rcseng.ac.uk/patient-care/surgical-staff-and-regulation/surgical-outcomes/> [Accessed].
- Rubin, J. & Chisnell, D. 2008. *How to plan, design, and conduct effective tests*, John Wiley & Sons Inc.
- Ryu, W. H. A., Mostafa, A. E., Dharampal, N., Sharlin, E., Kopp, G., Jacobs, W. B., . . . Sutherland, G. R. 2017. Design-Based Comparison of Spine Surgery Simulators: Optimizing Educational Features of Surgical Simulators. *World Neurosurgery*, 106, 870-877 e1.
- Sandars, J. & Cleary, T. J. 2011. Self-regulation theory: applications to medical education: AMEE Guide No. 58. *Medical Teacher*, 33, 875-86.
- Sandars, J. & Lafferty, N. 2010. Twelve Tips on usability testing to develop effective e-learning in medical education. *Medical Teacher*, 32, 956-960.
- Sandelowski, M. 1995. Sample size in qualitative research. *Research in Nursing and Health*, 18, 179-83.
- Sargeant, J., Lockyer, J., Mann, K., Holmboe, E., Silver, I., Armson, H., . . . Power, M. 2015. Facilitated Reflective Performance Feedback: Developing an Evidence- and Theory-Based Model That

- Builds Relationship, Explores Reactions and Content, and Coaches for Performance Change (R2C2). *Academic Medicine*, 90, 1698-706.
- Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., . . . Jinks, C. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant*, 52, 1893-1907.
- Schleiermacher, F. 1838. *Hermeneutik und Kritik mit besonderer Beziehung auf das Neue Testament*, Cambridge University Press.
- Schmidt, R. A. & Bjork, R. A. 1992. New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science*, 3, 207-218.
- Schoenfeld, G. O., Amdur, R. J., Morris, C. G., Li, J. G., Hinerman, R. W. & Mendenhall, W. M. 2008. Patterns of failure and toxicity after intensity-modulated radiotherapy for head and neck cancer. *International Journal of Radiation Oncology, Biology, Physics*, 71, 377-85.
- Seppenwoolde, Y., Assenholt, M. S., Georg, D., Nout, R., Tan, L. T., Rumpold, T., . . . Tanderup, K. 2019. Importance of training in external beam treatment planning for locally advanced cervix cancer: Report from the EMBRACE II dummy run. *Radiotherapy and Oncology*, 133, 149-155.
- Setyonugroho, W., Kennedy, K. M. & Kropmans, T. J. B. 2015. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Education and Counseling*, 98, 1482-1491.
- Sewell, J. L., Boscardin, C. K., Young, J. Q., Ten Cate, O. & O'sullivan, P. S. 2016. Measuring cognitive load during procedural skills training with colonoscopy as an exemplar. *Medical Education*, 50, 682-92.
- Sewell, J. L., Maggio, L. A., Ten Cate, O., Van Gog, T., Young, J. Q. & O'sullivan, P. S. 2019. Cognitive load theory for training health professionals in the workplace: A BEME review of studies among diverse professions: BEME Guide No. 53. *Medical Teacher*, 41, 256-270.
- Shariff, F., Hatala, R. & Regehr, G. 2020. Learning After the Simulation Is Over: The Role of Simulation in Supporting Ongoing Self-Regulated Learning in Practice. *Academic Medicine*, 95, 523-526.
- Sharma, R., Gordon, M., Dharamsi, S. & Gibbs, T. 2015. Systematic reviews in medical education: a practical approach: AMEE guide 94. *Medical Teacher*, 37, 108-24.
- Shute, V. J. 2008. Focus on formative feedback. *Review of Educational Research*, 78, 153-189.
- Sinha, P., Hogle, N. J. & Fowler, D. L. 2008. Do the laparoscopic skills of trainees deteriorate over time? *Surgical Endoscopy and Other Interventional Techniques*, 22, 2018-2025.
- Small, W., Jr., Mell, L. K., Anderson, P., Creutzberg, C., De Los Santos, J., Gaffney, D., . . . Mundt, A. J. 2008. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy in postoperative treatment of endometrial and cervical cancer. *International Journal of Radiation Oncology, Biology, Physics*, 71, 428-34.
- Stokes, D. E. 1997. *Pasteur's quadrant: Basic science and technological innovation*, Brookings Institution Press.
- Sturdza, A., Pötter, R., Fokdal, L. U., Haie-Meder, C., Tan, L. T., Mazeron, R., . . . Lindegaard, J. C. 2016. Image guided brachytherapy in locally advanced cervical cancer: Improved pelvic control and

- survival in RetroEMBRACE, a multicenter cohort study. *Radiotherapy and Oncology*, 120, 428-433.
- Sullivan, G. M., Simpson, D., Cook, D. A., Deiorio, N. M., Andolsek, K., Opas, L., . . . Yarris, L. M. 2014. Redefining Quality in Medical Education Research: A Consumer's View. *Journal of Graduate Medical Education*, 6, 424-9.
- Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257-285.
- Sweller, J., Van Merriënboer, J. J. G. & Paas, F. 2019. Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, 31, 261-292.
- Tai, P., Van Dyk, J., Battista, J., Yu, E., Stitt, L., Tonita, J., . . . Youssef, Y. 2002. Improving the consistency in cervical esophageal target volume definition by special training. *International Journal of Radiation Oncology*Biological*Physics*, 53, 766-774.
- Takiar, V., Fontanilla, H. P., Eifel, P. J., Jhingran, A., Kelly, P., Iyer, R. B., . . . Klopp, A. 2013. Anatomic distribution of fluorodeoxyglucose-avid para-aortic lymph nodes in patients with cervical cancer. *International Journal of Radiation Oncology, Biology, Physics*, 85, 1045-50.
- Tan, L. T. 2010. Advanced Radiotherapy (ART) - ART 01 - Image-Guided Brachytherapy for Cervix Cancer eLearning for Health.
- Tan, L. T., Potter, R., Sturdza, A., Fokdal, L., Haie-Meder, C., Schmid, M., . . . Tanderup, K. 2019a. Change in Patterns of Failure After Image-Guided Brachytherapy for Cervical Cancer: Analysis From the RetroEMBRACE Study. *International Journal of Radiation Oncology, Biology, Physics*, 104, 895-902.
- Tan, L. T., Tanderup, K., Kirisits, C., De Leeuw, A., Nout, R., Duke, S., . . . Potter, R. 2019b. Image-guided Adaptive Radiotherapy in Cervical Cancer. *Seminars in Radiation Oncology*, 29, 284-298.
- Tan, L. T., Tanderup, K., Kirisits, C., Mahantshetty, U., Swamidas, J., Jurgenliemk-Schulz, I., . . . Potter, R. 2020. Education and training for image-guided adaptive brachytherapy for cervix cancer-The (GEC)-ESTRO/EMBRACE perspective. *Brachytherapy*.
- Tanderup, K., Nesvacil, N., Kirchheiner, K., Serban, M., Spampinato, S., Jensen, N. B. K., . . . Potter, R. 2020. Evidence-Based Dose Planning Aims and Dose Prescription in Image-Guided Brachytherapy Combined With Radiochemotherapy in Locally Advanced Cervical Cancer. *Seminars in Radiation Oncology*, 30, 311-327.
- Tanderup, K., Nesvacil, N., Potter, R. & Kirisits, C. 2013. Uncertainties in image guided adaptive cervix cancer brachytherapy: impact on planning and prescription. *Radiotherapy and Oncology*, 107, 1-5.
- Tavakol, M. & Sandars, J. 2014. Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part I. *Medical Teacher*, 36, 746-756.
- Ten Cate, O. 2005. Entrustability of professional activities and competency-based training. *Medical Education*, 39, 1176-7.
- Ten Cate, O. T. 2013. Why receiving feedback collides with self determination. *Adv Health Sci Educ Theory Pract*, 18, 845-9.

- The Design-Based Research Collective 2003. Design-Based Research: An Emerging Paradigm for Educational Inquiry. *Educational Researcher*, 32, 5-8.
- The Mathworks Inc. 2018. MATLAB (R2018b). 9.7.0.1190202 (R2019b) ed. Natick, Massachusetts.
- The Mathworks Inc. 2019. MATLAB (R2019b). 9.7.0.1190202 (R2019b) ed. Natick, Massachusetts.
- The Quality Assurance Agency 2014. UK Quality Code for Higher Education.
- The Royal College of Surgeons of Edinburgh. 2021. *Intercollegiate Surgical Logbooks* [Online]. Available: <https://www.rcsed.ac.uk/professional-support-development-resources/learning-resources/intercollegiate-surgical-logbooks> [Accessed].
- U.S. General Services Administration Technology Transformation Services. *Usability.gov* : "How To & Tools" [Online]. Available: <https://www.usability.gov/how-to-and-tools/index.html> [Accessed 17/12/2020].
- University of Nottingham 2018. Nottingham Qualitative Methods Tutor (Online Course).
- Uttl, B., White, C. A. & Gonzalez, D. W. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22-42.
- Valentini, V., Boldrini, L., Damiani, A. & Muren, L. P. 2014. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. *Radiotherapy and Oncology*, 112, 317-20.
- Van De Ridder, J. M., Mcgaghie, W. C., Stokking, K. M. & Ten Cate, O. T. 2015. Variables that affect the process and outcome of feedback, relevant for medical training: a meta-review. *Medical Education*, 49, 658-73.
- Van De Ridder, J. M., Stokking, K. M., Mcgaghie, W. C. & Ten Cate, O. T. 2008. What is feedback in clinical education? *Medical Education*, 42, 189-97.
- Van De Steene, J., Linthout, N., De Mey, J., Vinh-Hung, V., Claassens, C., Noppen, M., . . . Storme, G. 2002. Definition of gross tumor volume in lung cancer: Inter-observer variability. *Radiotherapy and Oncology*, 62, 37-39.
- Van Den Akker, J. 1999. Principles and methods of development research. *Design approaches and tools in education and training*. Springer.
- Van Der Kleij, F. M., Feskens, R. C. W. & Eggen, T. J. H. M. 2015. Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes. *Review of Educational Research*, 85, 475-511.
- Van Der Vleuten, C. P. 1996. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract*, 1, 41-67.
- Van Gog, T., Paas, F. & Van Merriënboer, J. J. G. 2008. Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. *Learning and Instruction*, 18, 211-222.

- Van Merriënboer, J. J. G. & Sweller, J. 2010. Cognitive load theory in health professional education: design principles and strategies. *Medical Education*, 44, 85-93.
- Varpio, L., Paradis, E., Uijtdehaage, S. & Young, M. 2020. The Distinctions Between Theory, Theoretical Framework, and Conceptual Framework. *Academic Medicine*, 95, 989-994.
- Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.
- Vesper, J. 2014. *Developing expertise of those handling temperature-sensitive pharmaceutical products using e-learning*. PhD thesis, Murdoch University.
- Vickers, A. J., Bianco, F. J., Serio, A. M., Eastham, J. A., Schrag, D., Klein, E. A., . . . Scardino, P. T. 2007. The surgical learning curve for prostate cancer control after radical prostatectomy. *Journal of the National Cancer Institute*, 99, 1171-7.
- Vinod, S. K., Jameson, M. G., Min, M. & Holloway, L. C. 2016a. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiotherapy and Oncology*, 121, 169-179.
- Vinod, S. K., Lim, K., Bell, L., Veera, J., Ohanessian, L., Juresic, E., . . . Holloway, L. 2017. High-risk CTV delineation for cervix brachytherapy: Application of GEC-ESTRO guidelines in Australia and New Zealand. *Journal of Medical Imaging and Radiation Oncology*, 61, 133-140.
- Vinod, S. K., Min, M., Jameson, M. G. & Holloway, L. C. 2016b. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *Journal of Medical Imaging and Radiation Oncology*, 60, 393-406.
- Viswanathan, A. N., Erickson, B., Gaffney, D. K., Beriwal, S., Bhatia, S. K., Lee Burnett, O., 3rd, . . . Bosch, W. 2014. Comparison and consensus guidelines for delineation of clinical target volume for CT- and MR-based brachytherapy in locally advanced cervical cancer. *International Journal of Radiation Oncology, Biology, Physics*, 90, 320-8.
- Vygotsky, L. S. 1962. *Thought and language*, MIT press.
- Wang, F. & Hannafin, M. J. 2005. Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53, 5-23.
- Warfield, S. K., Zou, K. H. & Wells, W. M. 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23, 903-21.
- Weber, D. C., Poortmans, P. M., Hurkmans, C. W., Aird, E., Gulyban, A. & Fairchild, A. 2011. Quality assurance for prospective EORTC radiation oncology trials: the challenges of advanced technology in a multicenter international setting. *Radiotherapy and Oncology*, 100, 150-6.
- Weber, D. C., Tomsej, M., Melidis, C. & Hurkmans, C. W. 2012. QA makes a clinical trial stronger: evidence-based medicine in radiation therapy. *Radiotherapy and Oncology*, 105, 4-8.
- Webster, A., Appelt, A. L. & Eminowicz, G. 2020. Image-Guided Radiotherapy for Pelvic Cancers: A Review of Current Evidence and Clinical Utilisation. *Clinical Oncology (Royal College of Radiologists)*, 32, 805-816.
- Weiner, B. 1972. Attribution theory, achievement motivation, and the educational process. *Review of Educational Research*, 42, 203-215.

- Weiss, E., Richter, S., Krauss, T., Metzelthin, S. I., Hille, A., Pradier, O., . . . Hess, C. F. 2003. Conformal radiotherapy planning of cervix carcinoma: Differences in the delineation of the clinical target volume. A comparison between gynaecologic and radiation oncologists. *Radiotherapy and Oncology*, 67, 87-95.
- Wellington, J. 2015. *Educational research: Contemporary issues and practical approaches*, Bloomsbury Publishing.
- Whitehead, L. 2004. Enhancing the quality of hermeneutic research: decision trail. *Journal of Advanced Nursing*, 45, 512-8.
- Wong, G., Greenhalgh, T. & Pawson, R. 2010. Internet-based medical education: a realist review of what works, for whom and in what circumstances. *BMC Medical Education*, 10, 12.
- Yen, P. Y. & Bakken, S. 2012. Review of health information technology usability study methodologies. *Journal of the American Medical Informatics Association*, 19, 413-422.
- Yeung, A. R., Pugh, S. L., Klopp, A. H., Gil, K. M., Wenzel, L., Westin, S. N., . . . Kachnic, L. A. 2020. Improvement in Patient-Reported Outcomes With Intensity-Modulated Radiotherapy (RT) Compared With Standard RT: A Report From the NRG Oncology RTOG 1203 Study. *Journal of Clinical Oncology*, 38, 1685-1692.
- Young, J. Q., Van Merriënboer, J., Durning, S. & Ten Cate, O. 2014. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Medical Teacher*, 36, 371-84.
- Young, M., Thomas, A., Lubarsky, S., Ballard, T., Gordon, D., Gruppen, L. D., . . . Durning, S. J. 2018. Drawing Boundaries: The Difficulty in Defining Clinical Reasoning. *Academic Medicine*, 93, 990-995.
- Yudkowsky, R., Park, Y. S. & Downing, S. M. 2019. Introduction to assessment in the health professions. *Assessment in Health Professions Education*.
- Yudkowsky, R., Park, Y. S., Lineberry, M., Knox, A. & Ritter, E. M. 2015. Setting mastery learning standards. *Academic Medicine*, 90, 1495-500.
- Zeilefsky, M. J., Kollmeier, M., Cox, B., Fidaleo, A., Sperling, D., Pei, X., . . . Hunt, M. 2012. Improved clinical outcomes with high-dose image guided radiotherapy compared with non-IGRT for the treatment of clinically localized prostate cancer. *International Journal of Radiation Oncology, Biology, Physics*, 84, 125-9.
- Zendejas, B., Cook, D. A., Bingener, J., Huebner, M., Dunn, W. F., Sarr, M. G. & Farley, D. R. 2011. Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized controlled trial. *Annals of Surgery*, 254, 502-9; discussion 509-11.
- Zendejas, B., Wang, A. T., Brydges, R., Hamstra, S. J. & Cook, D. A. 2013. Cost: the missing outcome in simulation-based medical education research: a systematic review. *Surgery*, 153, 160-76.
- Zimmerman, B. J. & Shunk, D. H. 2011. Self-Regulated Learning and Performance. *Handbook of Self-Regulation of Learning and Performance*. Routledge.
- Ziv, A., Wolpe, P. R., Small, S. D. & Glick, S. 2003. Simulation-Based Medical Education: An Ethical Imperative. *Academic Medicine*, 78, 783-788.

APPENDICES

A Appendices

A.3 Chapter 3 Appendix

A.3.1 *Chapter 3 - additional tables and figures*

Appendix Table A.3-1 - Medical education textbooks forming the basis of the initial mapping of relevant educational theory. The initial search was conducted in 2017 - updated editions are shown where they have been published subsequently.

Textbook Title	Editor(s)	Date / Year
Essential skills for a medical teacher	Ronald Harden	2020 (3 rd Ed.) 2016 (2 nd Ed.)
Understanding medical education	Tim Swannick Kirsty Forrest Bridget O'Brien	2019 (3 rd Ed.) 2012 (2 nd Ed.)
Assessment in health professions education	Rachel Yudkowsky Yoon Soo Park Steven Downing	2019 (2 nd Ed.) 2013 (1 st Ed.)
Practical guide to the evaluation of clinical competence	Eric Holmboe Richard Hawkins Steven Durning	2018 (2 nd Ed.) 2008 (1 st Ed.)
ABC of teaching and learning in medicine	Peter Cantillon Diana Wood Sarah Yardley	2017 (3 rd Ed.)
Oxford textbook of medical education	Kieran Walsh	2016
Educational technologies in medical health sciences education	Susan Bridges Lap Ki Chan Cindy E Hmelo-Silver	2015
Researching medical education	Jennifer Cleland Steven Durning	2015
Medical education - theory and practice	Tim Dornan Karen Mann Albert Scherpbier John Spencer	2011
Medical education & training: from theory to delivery	Yvonne Carter Neil Jackson	2009

Appendix Table A.3-2 - Medical educational journals listed by impact factor (2019)

Medical education journal title	2019 impact factor*
Academic Medicine	5.4
Journal of Medical Internet Research	5.0
Medical Education	4.6
Medical Teacher	2.6
Advances in Health Sciences Education	2.5
Postgraduate Medicine	2.5
Journal of Surgical Education	2.2
Postgraduate Medical Journal	1.9
BMC Medical Education	1.8
Journal of Cancer Education	1.6
Journal of Continuing Education in the Health Professions	1.4
Journal of Postgraduate Medicine	1.2

* source: InCites Journal Citation Reports, Clarivate analytics (<https://jcr.clarivate.com/>)

A.5 Chapter 5 Appendix

Appendix Table A.5-1 - Collaborators on the EMBRACE-II EBRT contouring education and accreditation study

Institution	Collaborator(s)
Cambridge University Hospitals NHS Foundation Trust, UK	L.T. Tan
Department of Oncology, Aarhus University Hospital, Aarhus, Denmark	N.B.K. Jensen J.C. Lindegaard K. Tanderup
Department of Radiotherapy, Medical University of Vienna, Vienna, Austria	C. Kirisits R.C. Pötter T. Rumpold
Department of Radiation Oncology, University Medical Center, Utrecht, Netherlands	A.A.C. De Leeuw I.M. Jürgenliemk-Schulz
Department of Radiation Oncology, Erasmus Medical Center, Netherlands	R.A. Nout

A.6 Chapter 6 Appendix

Appendix Table A.6-1 - Collaborators on the EMBRACE-II brachytherapy contouring accreditation study

Institution	Collaborator(s)
Cambridge University Hospitals NHS Foundation Trust, UK	L.T. Tan
Department of Oncology, Aarhus University Hospital, Aarhus, Denmark	J.C. Lindegaard K. Tanderup
Department of Radiotherapy, Medical University of Vienna, Vienna, Austria	C. Kirisits R.C. Pötter M. Schmid A. Sturdza N. Nesvacil T. Rumpold
Department of Radiation Oncology, University Medical Center, Utrecht, Netherlands	A.A.C. De Leeuw I.M. Jürgenliemk-Schulz
Department of Radiation Oncology, Erasmus Medical Center, Netherlands	R.A. Nout
Department of Radiation Oncology, TATA memorial medical centre, Mumbai, India	U. Mahantshetty

Appendix Table A.6-2 - Jaccard Conformity Index (JCI) analysis per ROI. Presented 'Overall' (both cases) and per case

Case	ROI	Median JCI	Mean JCI	Interquartile range	Range
Overall	GTV	0.39	0.40	0.28 - 0.48	0.13 - 0.74
	HR-CTV	0.73	0.70	0.65 - 0.78	0.3 - 0.84
	IR-CTV	0.58	0.59	0.54 - 0.66	0.35 - 0.78
	Rectum	0.79	0.79	0.76 - 0.82	0.61 - 0.91
	Sigmoid	0.64	0.60	0.42 - 0.79	0.2 - 0.87
	Bladder	0.86	0.85	0.83 - 0.9	0.58 - 0.93
	Bowel	0.32	0.31	0.12 - 0.45	0 - 0.81
TATA1	GTV	0.29	0.30	0.21 - 0.37	0.17 - 0.62
	HR-CTV	0.73	0.71	0.68 - 0.79	0.3 - 0.82
	IR-CTV	0.60	0.59	0.54 - 0.65	0.35 - 0.78
	Rectum	0.80	0.81	0.78 - 0.83	0.73 - 0.91
	Sigmoid	0.79	0.78	0.75 - 0.84	0.35 - 0.87
	Bladder	0.90	0.88	0.86 - 0.91	0.58 - 0.93
	Bowel	0.41	0.43	0.34 - 0.51	0.15 - 0.77
TATA2	GTV	0.46	0.49	0.42 - 0.59	0.13 - 0.74
	HR-CTV	0.72	0.70	0.64 - 0.77	0.44 - 0.84
	IR-CTV	0.57	0.59	0.53 - 0.66	0.4 - 0.72
	Rectum	0.78	0.78	0.75 - 0.82	0.61 - 0.9

Sigmoid	0.43	0.44	0.3 - 0.57	0.2 - 0.71
Bladder	0.83	0.82	0.8 - 0.86	0.66 - 0.89
Bowel	0.12	0.18	0.02 - 0.24	0 - 0.81

Appendix Table A.6-3 - Descriptive data for participant volumes with reference contour volumes provided for comparison

Case	ROI	Reference volume	Median participant volume	Interquartile range	Ra
Case 1	GTV _{res}	13	30	19 - 44	4 -
	HR-CTV	66	64	57 - 72	35
	IR-CTV	139	161	132 - 196	81
	Rectum	53	49	38 - 56	32
	Sigmoid	59	49	40 - 57	20
	Bladder	183	177	171 - 182	15
	Bowel	78	116	55 - 232	7 -
Case 2	GTV _{res}	8	10	6 - 15	2 -
	HR-CTV	33	27	21 - 32	15
	IR-CTV	76	77	62 - 96	35
	Rectum	53	46	39 - 55	21
	Sigmoid	180	85	59 - 157	32
	Bladder	93	94	87 - 98	43
	Bowel	17	67	48 - 134	1 -

A.7 Chapter 7 Appendix

A.7.1 *Mini-Contour initial specification (abridged)*

1) Most of existing functionality in current flash tool converted to HTML5:

- Multiple images per exercise / 'case' (most commonly will be < 5)
- Need to be able to draw more than 1 contour per image at present
- Reset exercise & delete contours
- Zoom, contrast
- Learner submit contour and see answer (although it will loaded from database – see below)
- Instructions / user guide

2) Additional functionality:

- Save learner contour:
 - ID (ideally pull from moodle if user being directed from there)
 - Time / date
 - Learner contour co-ordinate data
- Set one or more 'answer' contours - using same interface, but storing the contour labelled as 'teacher' rather than learner or uploading contour with image but stored separately. This should be visible on learner submitting their contour / pressing 'answer' button
- Pull out data (JSON format) for above from database for analysis (which I will do in MATLAB)
- Be able to retrieve multiple learner contours, and teacher contour(s) for display during a presentation (doesn't have to be in real time, can just retrieve after learners submit their contours)
- User feedback button to collect user feedback / bug data
- Be able to create a new exercise / 'case' without having to hard code it
- User input:
 - Select 'contour' icon & mouse L-click to create a point
 - Some forgiveness around closing pixel of polygon (e.g. 5px radius)
 - Drag and drop contour points to move them (if possible)
 - Zoom – use zoom 'bar'
 - Once zoomed – navigate by drag and drop (providing not pressing on a contour point)

A.8 Chapter 8 Appendix

A.8.1 Chapter 8 - additional tables & figures

Appendix Table A.8-1 - Mini-Contour usability issues by category

Category	Issue	Severity (1 mild - 3 high)	Frequency (count)
Navigation / viewing	Tried zooming using mousewheel	1	3
	Viewing window not maximised	1	2
	Contrast not windowing	1	1
	Overly sensitive image scroll [mouse trackpad]	1	1
	Inadvertently scrolled slices	1	1
	Tried to use keyboard shortcut from previous software [up & down arrows]	1	1
	Difficulty finding back button	1	1
	Selection of 'pan' unclear	1	1
Exercise content	Searched for guidance (ATLAS) for a long time	3	1
	Contoured on wrong side	2	2
	Insufficient image slices in exercise	2	1
	Insufficient knowledge - submitted without contour	2	1
	No right / left labels (contoured on correct side)	1	1
	Poor image resolution	1	1
	Exercise instructions unclear to user	1	1
Drawing	Inadvertent point	2	6
	Didn't fully close contour	1	1
	Tries to click & drag to draw	1	1
	Re-names contours unnecessarily	1	1
	Unclear if draw is selected	1	1
Editing	Difficulty adding points	2	3
	Can't undo action	2	1
	Can't see contour detail due to colourwash	1	1
	Tried to move point but closed contour	1	1
Submitting	Went in and out of exercise as thought it saved contour - had to repeat	2	1
Reviewing feedback	Key structure not shown as a learning zone	2	2
	Couldn't zoom out in feedback view	2	1
	Difficulty relating learning zone feedback to image	1	6
	Scrolled off feedback accidentally	1	3
	Difficulty identifying gold standard	1	2
	Comment zone not visible	1	1
	Thought 'comment' was a user note	1	1
	Toggle contour visibility not intuitive	1	1

A.8.2 Usability study recruitment email

SUBJECT LINE: Invitation to take part in the “Mini-Contour” user experience study

Hello,

My name is Simon Duke and I’m helping to run a study evaluating a new online learning tool for radiotherapy contouring – “Mini-Contour”. In order to improve the tool, we’re looking for people who may be interested in trying out the tool and giving feedback after using it.

You will receive £20 to participate. The study has ethical approval from Nottingham University School of Medicine.

What will I be doing in a usability study?

You will fill in a questionnaire. You will then be asked to do several short tasks using the learning tool. You will also be asked questions about your experience and perceptions of the website.

How long is a session?

One and a half hours.

When and where?

We can arrange a time and date that is convenient to you. You can participate online (over a web-conferenced call) or in person at Cambridge University Hospitals.

Interested in participating?

Please e-mail me at simon.duke@nhs.net. I’ll be in touch to give you some more details and ask you some questions to help us determine if you qualify for the study.

If you have any questions, please contact me at simon.duke@nhs.net.

Thank you for interest,

Dr Simon Duke

Specialist Registrar (ST7) in Clinical Oncology
Senior Clinical Research Fellow – Radiotherapy – Cambridge University Hospitals
Doctoral Research Degree Candidate – University of Nottingham School of Medicine

A.8.3 *Usability study test script*

Mini-Contour Usability Study - Test Script v1.0

*** Please note this script has been modified from the **un**-copyrighted materials available at www.usability.gov ** (U.S. General Services Administration Technology Transformation Services)*

Welcome and Purpose

Thank you so much for participating today. I want to give you a little information about what you will be looking at, and give you time to ask any questions you might have before we get started.

Today we are asking you to evaluate a radiotherapy learning tool and to complete a set of tasks. Our goal is to see how easy or difficult you find the tool to use.

Test Facilitator's Role

I am here to listen to your reactions and comments. During the practical session, I would like you to think aloud as you work to complete the tasks. I will not be able to offer any suggestions or hints, but from time to time, I may ask you to clarify what you have said or ask you for information on what you were looking for or what you expect to have happen.

Test Participant's Role

- Today I am going to be asking you to complete some exercises using the tool and tell me how easy or difficult it was. These activities are all about how easy we have made it for people to use the learning tool.
- There is no right or wrong answer. If you have any questions, comments or areas of confusion while you are working, please let me know.
- If you ever feel that you are lost or cannot complete a task with the information that you have been given, please let me know. I will ask you what you might do in a real-world setting and then either put you on the right track or let you continue. I may move you on to the next scenario or task.
- As you use the tool, please do so as you would at home or work. I would ask that you to try work through the tasks based on what you see on screen, but if you reach a point where you are not sure where or how to find something, please feel free to review the tutorial provided.
- We will be recording this session for reference if needed. We are capturing your face, your voice and what you see on the screen. Your name will not be associated or reported with data or findings from this evaluation.
- I may ask you other questions as we go and we will have some questions at the end.

Do you have any questions before we begin?

Example moderator questions:

- Talk me through the words, links, pictures and icons here ...
- What additional information would you want to see?
- What are you trying to do now?
- How do you think you would solve this problem? / How do you think that would work?
- How easy or difficult was that for you?

A.8.4 *Usability study pre-task questionnaire*

*Rules for skipping questions are in **green***

1) Demographic details:

- Study user ID
- Age
- Gender
- Country of residence
- Are you (choose one):
 - In training in clinical / radiation oncology in the UK
 - In training in clinical / radiation oncology outside of the UK
 - A certified practitioner in clinical / radiation oncology

➔ *If selected option (i), skip to question 2.*

➔ *If selected option (ii), skip to question 4.*

➔ *If selected option (iii), skip to question 5.*

To participants who are in training

2) What is your Clinical Oncology stage of training (if applicable):

- ST3
- ST4
- ST5
- ST6
- ST7 +

3) What is your Clinical Oncology training exam progress (if applicable)

- Pre FRCR part 1
- Post FRCR part 1
- Post Final FRCR part 2a
- Post Final FRCR part 2b

- 4) How many years have you been in specialist radiation oncology / clinical oncology training?
- a. Less than 1
 - b. 1-2
 - c. 2-4
 - d. 4 or more

➔ *Skip to question 7*

To participants who are consultants / certified practitioners

- 5) How long have you been qualified as a consultant or certified radiation oncologist / clinical oncologist?
- a. Less than 1 year
 - b. 1-2 years
 - c. 2-5 years
 - d. 5-10 years
 - e. > 10 years (please specify)

- 6) What are your specialist tumour sites?

[Open ended]

Delineation training questions – to all groups

- 7) What training in delineation have you had (select all that apply)?
- a. No formal training
 - b. Trained by a colleague in same department
 - c. Visited another department
 - d. Attended live teaching course
 - e. Participated in online / e-learning course
 - f. Participated in external audit / quality assurance (e.g. for a clinical trial)

- 8) Have you used a software tool for practicing delineation before?

No – *finish survey*

Yes - If so, which?

- a. Educase / FALCON
- b. ProKnow
- c. Aquilab
- d. Mini-Contour
- e. Addenbrooke's Contouring Tool (EMBRACE group)
- f. Other (please specify)

9) Describe your experiences of using the software tool(s) above

[Open ended]

10) How long do you estimate it took to complete a practice exercise / case using the software tool above (answer for each)?

- a. Less than 10 minutes
- b. 10-30 minutes
- c. 30-60 minutes
- d. 1-2 hours
- e. More than 2 hours

A.8.5 Usability study post-task questionnaire & interview

1) Demographic details:

- a. Study user ID

2) Unified theory of acceptance and Use of Technology (UTAUT) questionnaire:

Modifications from standard questionnaire:

- “system” replaced with “tool”
- Other modifications highlighted in yellow

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
I would find the tool useful in my job	1	2	3	4	5
Using the tool enables me to accomplish tasks more quickly	1	2	3	4	5
Using the tool increases my productivity	1	2	3	4	5
If I use the tool, I will increase my chances of progressing in my career	1	2	3	4	5
My interaction with the tool is clear and understandable	1	2	3	4	5
It would be easy for me to become skilful at using the tool	1	2	3	4	5
I would find the tool easy to use	1	2	3	4	5
Learning to operate the tool is easy for me	1	2	3	4	5
Using the tool is a good idea	1	2	3	4	5
The tool makes learning more interesting	1	2	3	4	5
Working with the tool is fun	1	2	3	4	5
I like working with the tool	1	2	3	4	5
People who influence my behaviour think that I should use the tool	1	2	3	4	5
People who are important to me think that I should use the tool	1	2	3	4	5
The senior management of my organisation have been helpful in the use of the tool	1	2	3	4	5
In general, my organisation has supported the use of the tool	1	2	3	4	5

I have the resources necessary to use the tool	1	2	3	4	5
I have the knowledge necessary to use the tool	1	2	3	4	5
The tool is not compatible with other systems I use	1	2	3	4	5
A specific person (or group) is available for assistance with difficulties with the tool	1	2	3	4	5
I could complete a task using the tool:					
If there was no-one around to tell me what to do as I go	1	2	3	4	5
If I could call someone for help If I got stuck	1	2	3	4	5
If I had a lot of time to complete the job for which the software was provided	1	2	3	4	5
If I had just the built-in help facility for assistance	1	2	3	4	5
I feel apprehensive about using the tool	1	2	3	4	5
It scares me to think that I could lose a lot of information by hitting the wrong key	1	2	3	4	5
I hesitate to use the tool for fear of making mistakes I cannot correct	1	2	3	4	5
The tool is somewhat intimidating to me	1	2	3	4	5
I intend to use the tool in the next 12 months	1	2	3	4	5
I predict I would use the tool in the next 6 months	1	2	3	4	5
I plan to use the tool in the next 6 months	1	2	3	4	5
[Free text comment box here]					

Interview questions

- 3) How similar do you think Mini-Contour is to radiotherapy delineation / contouring in real life? Please give a percentage on a scale from 0 to 100%.

- 4) You rated the tool on how similar it is to delineation in real life ... Can you tell me about the reasons for the ratings you gave?
- 5) What do you think about the design (the look and feel) of the tool?
- 6) Can you tell me about your experience of drawing and editing your contours?
- 7) What were your reactions to the reference contour feedback?
- 8) What were your reactions to the learning zone feedback?
- 9) Do you think the learning zone feedback was valid? If so, how often? When wasn't it valid?
- 10) What are your thoughts about how this tool could be improved?
 - What do you think should be our top priority (for improvement)?

A.9 Chapter 9 Appendix

A.9.1 Chapter 9 tables & figures

Appendix Table A.9-1 - Collaborators for Mini-Contour pilot studies

Name	Institute	Role
University of Nottingham	Gill Doody	Chief investigator & PhD supervisor
	Rakesh Patel	Co-investigator & PhD supervisor
	Heather Wharrad	Co-investigator & PhD supervisor
Cambridge University Hospitals (NHS)	Li-Tee Tan	Co-investigator & PhD supervisor
University College London Hospitals (NHS)	Gemma Eminowicz (GE)	Local faculty
	Asma Sarwar	Local faculty
Imperial College Hospitals (NHS)	Ed Wong	Local faculty
Erasmus Medical Center, The Netherlands	Remi Nout (RN)	Local faculty
Medical Center of Vienna, Austria	Max Schmidt	Local faculty
	Alina Sturdza	

Appendix Table A.9-2 - Content analysis of UK trainees' comments and suggestions for Mini-Contour

Response theme	Frequency
Improve drawing / editing tools	29
General positive comment	27
Expand to other tumour sites	10
Reduce learning zone stringency	7
Expand number of exercises	5
Improve viewing tools (including contrast)	5
Place in training - early	4
Positive re: learning zone feedback	2
Provide more image slices	2
Difficult for inexperienced trainees	2
3D viewing	1
Acknowledge acceptable variation in feedback	1
Allows contour comparison with others	1
Enable personal notes	1
Individualised pacing helpful	1
Positive - anatomy feedback	1
Practice fusion imaging modalities	1
Reassuring seeing group variability	1
Shorten retention interval	1
Standardises contouring	1

Appendix Table A.9-3 - Correlation of UK trainees' cervix cancer experience with their ranked performance on the first relevant learning exercise ("N/S" = $p > 0.05$ unless specified)

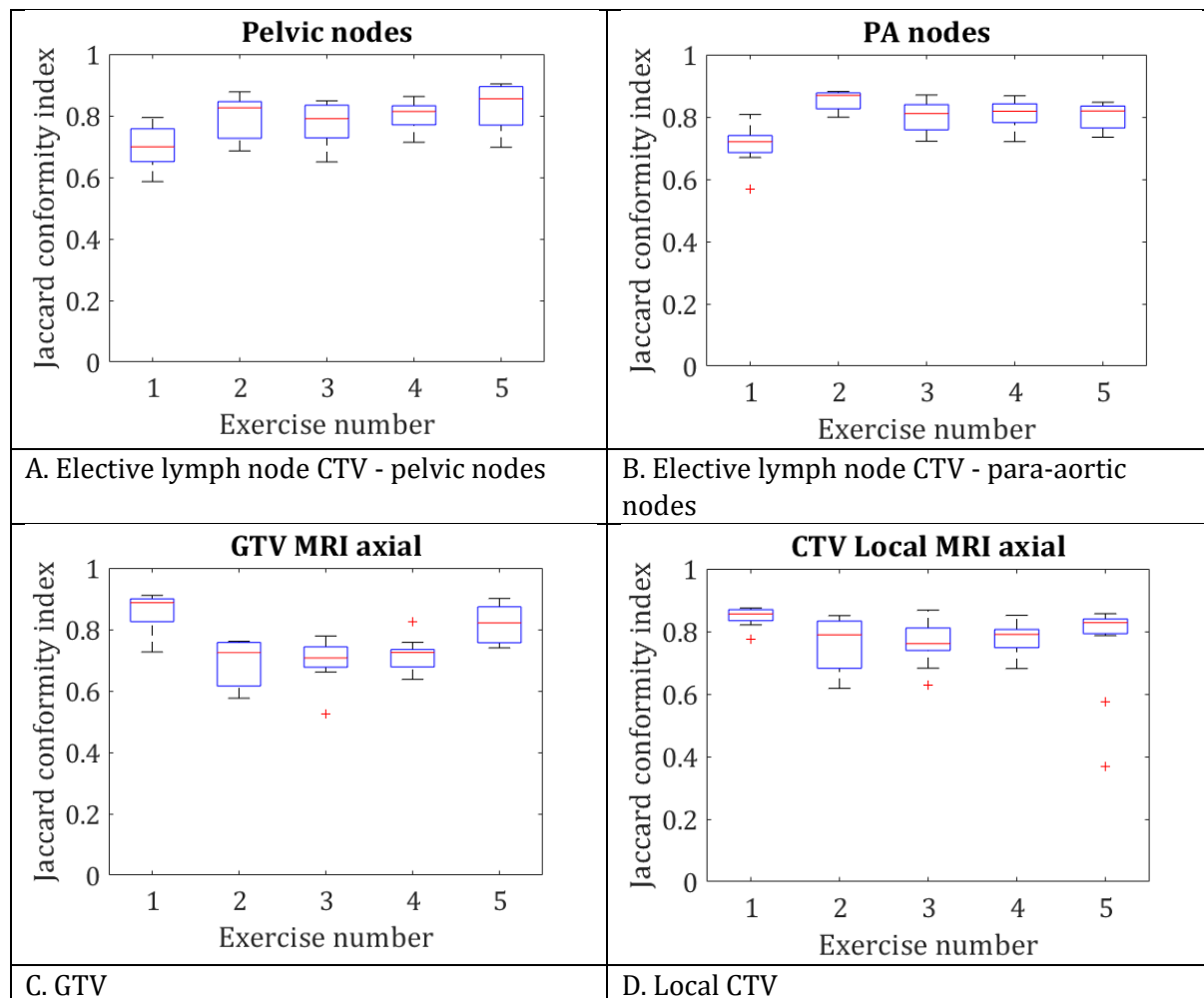
Region of interest	Correlation of cervix experience with initial performance by learning zones (Spearman's rho)	Correlation of cervix experience with initial performance by Jaccard conformity index (Spearman's rho)
GTV (on MRI)	0.17 (N/S)	0.27 ($p = 0.04$; N/S accounting for multiple testing)
Parametrium (Local CTV)	0.34 ($p = 0.007$)	0.21 (N/S)
CTV-E (Pelvic)	0.27 ($p = 0.04$; N/S accounting for multiple testing)	0.21 (N/S)
CTV-E (Para-aortic)	0.26 (N/S)	-0.05 (N/S)

Appendix Table A.9-4 - Analysis of the sensitivity of selected learning zones to variations in stringency in the UK trainees cohort

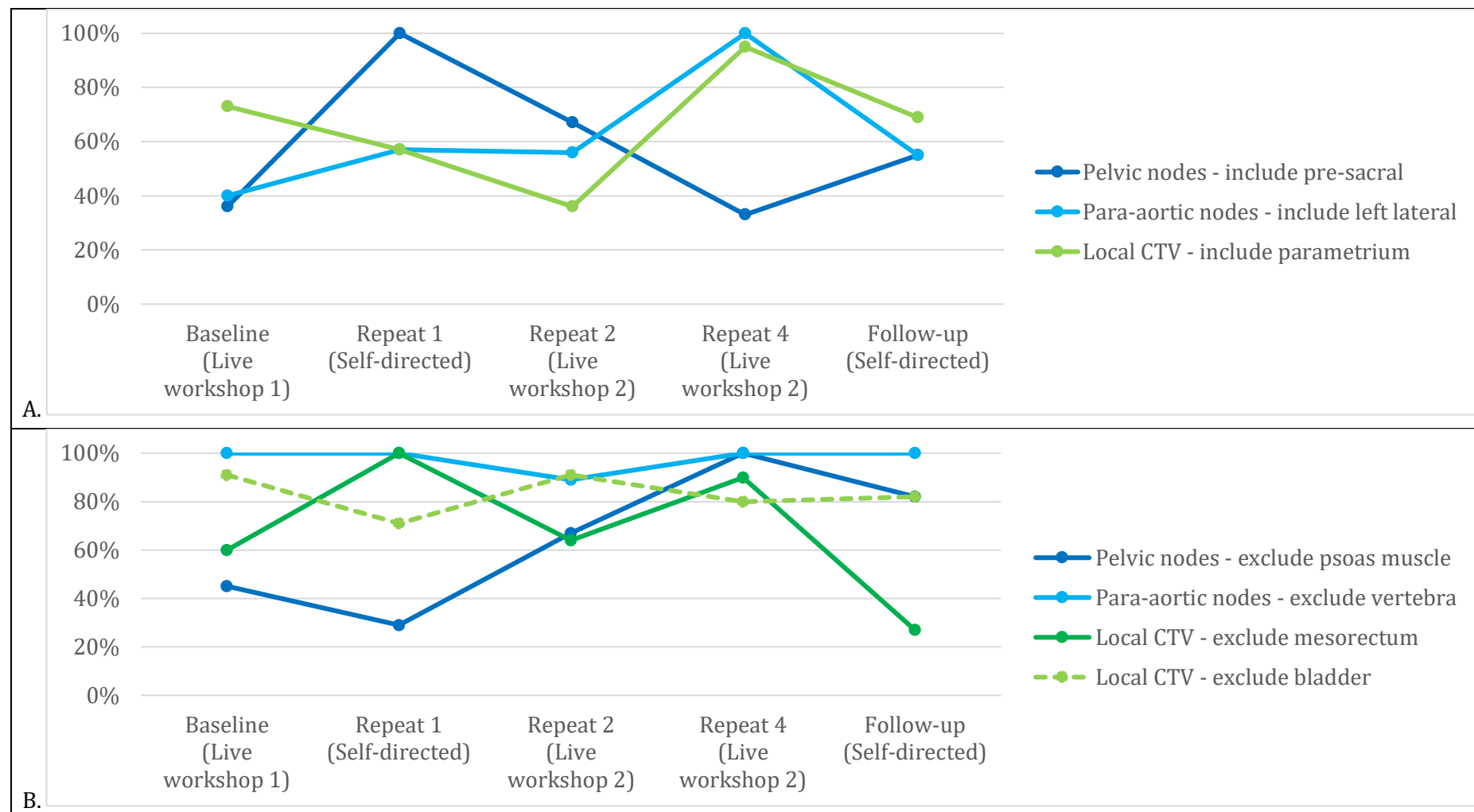
Learning zone	Acceptable Overlap threshold (least to most stringent)	Success rate - exercise 1	Success rate - exercise 2	Success rate - follow-up
Pelvic Nodes - include pre-sacral region	80%	62%	97%	50%
	90%	57%	97%	42%
	95%	54%	96%	39%
	98%	52%	96%	36%
	100%	41%	93%	31%
PA Nodes - include left lateral nodes	80%	62%	97%	35%
	90%	60%	94%	35%
	95%	55%	94%	35%
	98%	51%	91%	35%
	100%	48%	91%	32%
Local CTV - include parametrium	80%	76%	97%	72%
	90%	64%	94%	72%
	95%	55%	91%	72%
	98%	49%	88%	72%
	100%	43%	80%	69%
Pelvic Nodes - exclude bowel	20%	100%	99%	100%
	10%	100%	97%	100%
	5%	100%	96%	97%
	2%	100%	94%	97%
	0%	100%	82%	97%
Pelvic Nodes - exclude psoas	20%	100%	100%	94%
	10%	83%	100%	94%
	5%	83%	85%	83%
	2%	81%	76%	83%
	0%	73%	63%	81%
PA Nodes - exclude vertebral body	20%	100%	100%	100%
	10%	100%	100%	100%
	5%	100%	100%	100%
	2%	100%	100%	100%
	0%	94%	99%	94%
Local CTV - exclude bladder	20%	100%	100%	100%
	10%	100%	100%	100%
	5%	100%	97%	100%
	2%	97%	82%	100%
	0%	36%	26%	84%
Local CTV - exclude mesorectum	20%	97%	97%	94%
	10%	91%	97%	91%
	5%	90%	97%	72%
	2%	85%	92%	59%
	0%	64%	43%	25%

Appendix Table A.9-5 - Content analysis of international trainees' examples of when they learned from feedback

Feedback	Freq
Lateral PA nodes	12
Non-specific	7
Incorporate clinical examination findings	3
Obturator nodes	3
Parametrium	3
Cervix anatomy	2
Faculty contours helpful (general)	2
Lymph nodes (general)	1
MRI radiologic anatomy (general)	1
Organs at risk (general)	1
Total	35



Appendix Figure A.9-1 - International trainees' performance by Jaccard conformity index (JCI) over the 5 repeats of the four repeated exercise themes - **most engaged trainees only (n=11)**



Appendix Figure A.9-2 - International trainees' performance on 'include' (A) and 'exclude' (B) learning zones repeated over the course of the programme - most engaged trainees only (n=11)

Appendix Table A.9-6 - Content analysis of EMBRACE clinicians' comments regarding Mini-Contour fidelity

Response theme	Frequency
Lacking specific drawing / editing functionality	8
No 3D viewing	6
Lacking other viewing tools / quality	5
Limited clinical information	3
Positive comment re: rapid practice	2
Generally positive comment	2
Limited number of slices provided	2
Adaptations bypass lack of fidelity	3
Similar to real life	1
Lack multiple image series and modalities	1
Different workflow c.f. real life	1
Single slice contouring	1
Idealised image quality	1

Appendix Table A.9-7 - Content analysis of EMBRACE-II clinicians' general comments and suggestions

Response theme	Frequency
General positive comment	12
Improve drawing / editing tools	3
Enables rapid practice	3
Multiple reference contours	2
Further practice over time	3
Improve viewing tools	1
Positive comment re: learning zone feedback	2
General comment - luke warm	1
Suitable for trainees	1
Reduce learning zone stringency	1
Improve clinical information	1
More detailed learning zone feedback	1
More group discussion	1

Appendix Table A.9-8 - Correlation of confidence and performance in EMBRACE-II participants

Type	Exercise	Performance assessment	rho	p
EBRT	Elective lymph nodes - Pelvic	Learning Zones	-0.38	0.010
		JCI	-0.39	0.009
	Elective lymph nodes - Para-aortic	Learning Zones	-0.06	0.706
		JCI	-0.33	0.033
	Local CTV - Vagina	Learning Zones	-0.03	0.825
		JCI	0.04	0.787
Brachy-therapy	GTVres	Learning Zones	-0.17	0.271
		JCI	-0.07	0.644
	HR-CTV	Learning Zones	-0.14	0.363
		JCI	-0.08	0.620
	IR-CTV (Axial)	Learning Zones	0.05	0.784
		JCI	0.36	0.028
	IR-CTV (Sagittal)	Learning Zones	-0.07	0.675
		JCI	-0.04	0.830

Appendix Table A.9-9 - Comparison between the performance of principal investigators ("PIs") and other clinicians ("non-PIs")

Exercise	median PI JCI	median non-PI JCI	p	mean PI learning zone score	mean non-PI learning zone score	p
EBRT - Elective lymph node CTV - Pelvic nodes	0.84	0.83	0.09	0.92	0.84	0.13
EBRT - Elective lymph node CTV - PA nodes	0.86	0.84	0.06	0.97	0.90	0.07
EBRT - Local CTV - Parametrium	0.87	0.89	0.37	0.90	0.90	0.67
EBRT - Local CTV - Vagina	0.80	0.73	0.35	0.70	0.67	0.63
Brachytherapy - GTVres	0.49	0.53	0.02	0.46	0.79	0.02
Brachytherapy - HR-CTV	0.77	0.77	0.87	0.69	0.74	0.50
Brachytherapy - IR-CTV (axial)	0.79	0.77	0.44	0.63	0.63	1.00
Brachytherapy - IR-CTV (sagittal)	0.79	0.79	0.51	0.63	0.55	0.22

A.9.2 *Pre-task questionnaire*

Rules for skipping questions are in **green**

1) Demographic details:

- a. Study user ID:
- b. Are you (choose one):
 - i. In training in clinical / radiation oncology in the UK
 - ii. In training in clinical / radiation oncology outside of the UK
 - iii. A certified practitioner in clinical / radiation oncology

➔ *If selected option (i), skip to question 2.*

➔ *If selected option (i), skip to question 4.*

➔ *If selected option (ii), skip to question 6.*

To participants who are in training:

- 2) What is your Clinical Oncology stage of training (if applicable)?
 - a. ST3
 - b. ST4
 - c. ST5
 - d. ST6
 - e. ST7 +
- 3) What is your Clinical Oncology training exam progress (if applicable)?
 - a. Pre FRCR part 1
 - b. Post FRCR part 1
 - c. Post Final FRCR part 2a
 - d. Post Final FRCR part 2b
- 4) What is your clinical experience in cervix cancer?
 - a. None
 - b. 1-4 months
 - c. 4-8 months
 - d. 8-12 months
 - e. 1-2 years
 - f. > 2 years (specify no. of years)

➔ *Skip to question 8*

To participants who are consultants / certified practitioners

- 5) How long have you personally treated cervix cancer patients for?
 - a. 0 – 6 months
 - b. 6months - 1 year
 - c. 1-2 years
 - d. 2-5 years
 - e. 5-10 years
 - f. > 10 years (please specify)

- 6) How many cervix cancer patients do you personally treat each year?
- 0
 - 1-10
 - 11-25
 - 26-50
 - >50
- 7) Which guidelines do you use for contouring IMRT for cervix cancer (select all that apply)?
- GYN IMRT consortium (Lim et al., Int J Radiat Oncol Biol Phys. 2011 Feb 1;79(2):348-55)
 - Taylor pelvic IMRT guidelines (Int J Radiat Oncol Biol Phys. 2005 Dec 1;63(5):1604-12)
 - National guidelines (please specify in the below box)
 - Trial protocol (please specify in the below box)
 - Local guidelines
 - No guidelines
 - Other (please specify in the below box)

Cervix cancer questions – to all groups

- 8) What training in delineation for cervix cancer IMRT have you had (select all that apply)?
- Have not used IMRT for cervical cancer
 - No formal training
 - Trained by a colleague in same department
 - Visited another department
 - Attended live teaching course
 - Participated in online / e-learning course
 - Participated in external audit / quality assurance (e.g. for a clinical trial)
- ➔ *If selected option (a), go to question 11.*
- 9) Do you use the concept of an Internal Target Volume (ITV) when contouring IMRT for cervical cancer?
- No
 - No, but allow for when contouring CTV
 - No, but allow for in PTV margin
 - Yes – draw as separate ROI
 - Other (please specify)
- 10) What imaging modalities do you routinely use for contouring for IMRT for cervix cancer (select all that apply)?
- Planning CT only
 - Planning CT with diagnostic MRI as a reference ('side-by-side')
 - Planning CT fused to diagnostic MRI
 - Planning CT fused to MRI in treatment position
 - PET-CT as a reference ('side-by-side')
 - PET-CT fused to either MRI or CT
- 11) How confident are you at contouring the following structures for cervical cancer IMRT?

	Not at all confident				Very confident
Initial Tumour GTV on MRI	1	2	3	4	5

Nodal GTV	1	2	3	4	5
Local CTV - Parametrium	1	2	3	4	5
Local CTV - Vagina	1	2	3	4	5
Elective nodal CTV – pelvic nodes	1	2	3	4	5
Elective nodal CTV – para-aortic nodes	1	2	3	4	5
ITV	1	2	3	4	5

12) How confident are you at contouring the following structures for cervical cancer IMRT?

	1 Not at all confident	2	3	4	5 Very confident
Bladder	1	2	3	4	5
Rectum	1	2	3	4	5
Sigmoid	1	2	3	4	5
Bowel – “bowel bag”	1	2	3	4	5
Bowel – “bowel loops”	1	2	3	4	5
Bowel – “peritoneal cavity”	1	2	3	4	5

Additional questions for EMBRACE-II clinicians

13) How confident are you at contouring the following structures for cervical cancer brachytherapy?

	Not at all confident				Very confident
Residual GTV	1	2	3	4	5
High-Risk CTV	1	2	3	4	5
Intermediate-Risk CTV	1	2	3	4	5
Bladder	1	2	3	4	5
Rectum	1	2	3	4	5
Sigmoid	1	2	3	4	5
Bowel	1	2	3	4	5

A.9.3 Quick contouring guide for trainees (cervix cancer)

ROI	Description
GTV-T (primary)	Combine findings from MRI (high signal on T2-weighted), CT and examination under anaesthetic (EUA)
CTV 'Local'	Includes: Gross tumour + Cervix Uterus Parametria [guidelines inconsistent] Vagina - upper half, or ≥ 2 cm below tumour [guidelines inconsistent] [some guidelines include ovaries / proximal utero-sacral ligaments]
CTV 'Elective' / 'Elective Nodal' Uninvolved lymph node regions General: include all visible nodes / lymphoceles	Common Iliac nodes: expand vessels by 7mm and exclude muscle / bone. Join with 10mm strip across anterior sacrum. Superior border generally top of L5 / level of aortic bifurcation
	Internal / External Iliac: expand vessels by 7mm and exclude muscle / bone. Join with 10-18mm strip parallel to the pelvic sidewall. Stop contouring external iliac when goes into inguinal compartment
	Pre-sacral: 10mm strip anterior to sacrum down to inferior border of S2
	Obturator: 10-18mm strip medial to pelvic sidewall (do not include muscle or bone)
	Inguinal: Only include if lower 1/3 vagina involved. Minimum 7-10mm around vessels. Bordered by pectineus, iliopsoas and sartorius muscles.
	Para-aortic nodes: expand vessels by 7mm then adjust manually to include left para-aortic area. Exclude muscle and bone. Top of L1/L2 to bottom L4.
GTV-N (node)	Pathologically involved lymph nodes PET-CT may help distinguish pathological nodes
CTV-N (involved node)	GTV-N + margin (often 5mm)
Internal Target Volume ("ITV")	Still under development. EMBRACE-II trial uses: Outline 'CTV Local' on CT and MR and merge Add additional 'safety margin' to region gross tumour / cervix Individualise (enlarge) this volume to accommodation direction of potential movement Then add elective nodal CTV

Ref:

International Consensus Guidelines - Lim, IJROBP 2011; 79(2)348-355

INTERLACE Trial Protocol

EMBRACE-II Trial Protocol - www.embracestudy.dk

Japanese Consensus Guidelines - Toita, Jpn J Clin Oncol 2011;41(9)1119-1126

A.9.4 *Post-task questionnaire*

Collected within the tool re: specific learning zones: do you agree with the location and feedback for this learning zone? (5-point scale – “Not at all” to “Very much”). Free-text comment.

1) Demographic details:

- Study user ID

2) Did you share a laptop or computer for the delineation exercises?

- Yes
- No
- If yes, with which other study ID(s) [Free text]

3) System Usability Scale [scores in grey were not visible to the participants]:

Regarding the Mini-Contour Tool:	Strongly disagree				Strongly agree
I think that I would like to use this tool frequently	1	2	3	4	5
I found the tool unnecessarily complex	5	4	3	2	1
I thought the tool was easy to use	1	2	3	4	5
I think that I would need the support of a technical person to be able to use this tool	5	4	3	2	1
I found the various functions in this tool were well integrated	1	2	3	4	5
I thought there was too much inconsistency in this tool	5	4	3	2	1
I would imagine that most people would learn to use this tool very quickly	1	2	3	4	5
I found the tool very cumbersome to use	5	4	3	2	1
I felt very confident using the tool	1	2	3	4	5
I needed a lot of instruction / help before I could get going with this tool	5	4	3	2	1
Comments:					

4) How similar do you think Mini-Contour is to radiotherapy delineation / contouring in real life? Please give a percentage on a scale from 0 to 100%.

5) Please give a reason for your answer above [added via protocol amendment: international trainees & EMBRACE group only]

6) Regarding the Mini-Contour tool and the delineation workshop:

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The Mini-Contour tool is useful for learning radiotherapy delineation	1	2	3	4	5
I enjoyed learning using the Mini-Contour tool	1	2	3	4	5
The delineation workshop(s) improved my confidence in cervix cancer IMRT delineation	1	2	3	4	5
I would be interested in future delineation practice using this tool	1	2	3	4	5
Comments:					

- 7) What are your thoughts about how this tool could be improved? [Free text comment]
- 8) Repeat questions on confidence contouring cervix cancer EBRT targets (*see above*)
- 9) Repeat questions on confidence contouring cervix cancer EBRT organs at risk (*see above*)
- 10) Would you be willing to be contacted about future studies involving this learning approach? (yes/no)

Additional questions for international trainees in longitudinal programme and EMBRACE-II clinicians

- 11) Concerning the learning zone feedback in general:

	Not at all		Somewhat		Very much
Did you agree with the location and feedback of 'learning zones' in general?	1	2	3	4	5
How useful was the 'learning zone' location and feedback in general	1	2	3	4	5
How clinically relevant were the 'learning zones' in general?	1	2	3	4	5

- 12) Can you give an example of when you learned something from a 'learning zone?'
- 13) Can you give an example of when you disagreed with a 'learning zone?'

Additional questions for EMBRACE-II clinicians

- 14) Repeat confidence in brachytherapy target volumes (see above)
- 15) Repeat confidence in brachytherapy organ at risk volumes (see above)